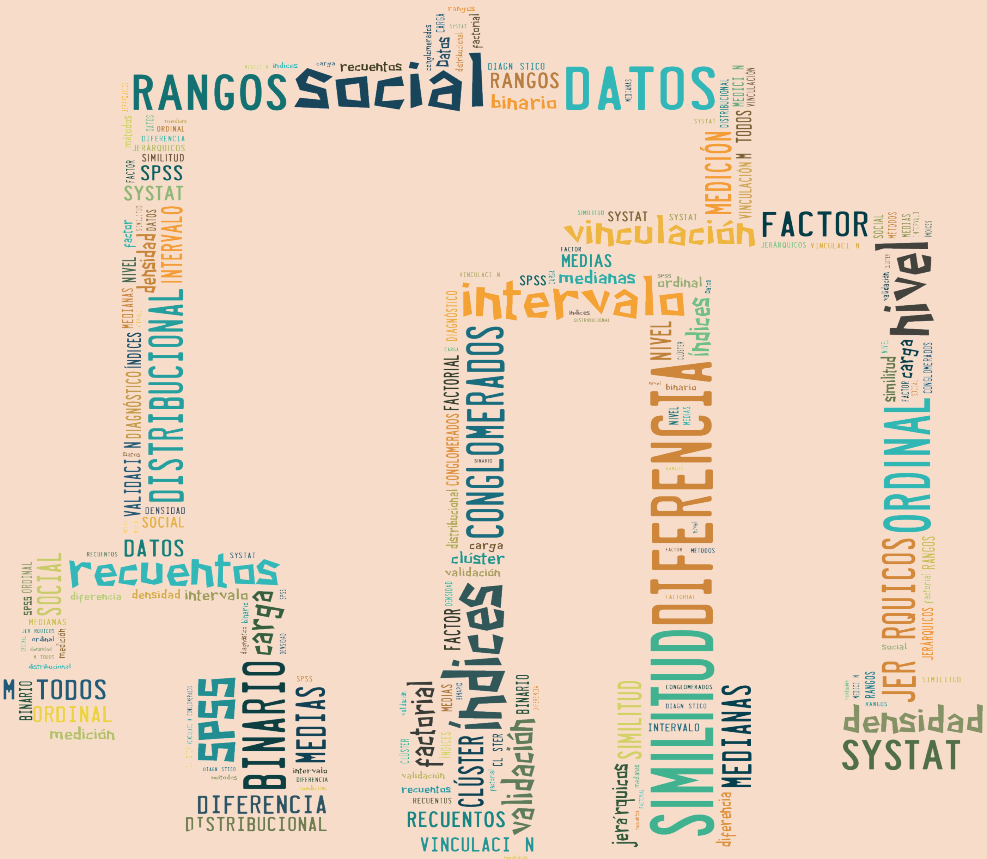
Antonio Alaminos Chica
Francisco José Francés García
Clemente Penalva Verdú
Óscar Antonio Santacreu Fernández

Análisis multivariante para las Ciencias Sociales I

Índices de distancia, conglomerados y análisis factorial



PYDLOS Ediciones

ANÁLISIS MULTIVARIANTE PARA LAS CIENCIAS SOCIALES I

Índices de distancia, conglomerados y análisis factorial

ANTONIO ALAMINOS CHICA
FRANCISCO FRANCÉS GARCÍA
CLEMENTE PENALVA VERDÚ
ÓSCAR SANTACREU FERNÁNDEZ

ANÁLISIS MULTIVARIANTE PARA LAS CIENCIAS SOCIALES I

ÍNDICES DE DISTANCIA, CONGLOMERADOS Y ANÁLISIS FACTORIAL

ANÁLISIS MULTIVARIANTE PARA LAS CIENCIAS SOCIALES I.
Índices de distancia, conglomerados y análisis factorial

Antonio Alaminos Chica
Francisco José Francés García
Clemente Penalva García
Óscar Antonio Santacreu Fernández

ISBN: 978-9978-14-315-5

Diseño Portada: Óscar Santacreu
Diagramación: Patricia Barbero
Corrección de estilo: María Eugenia Estrella
Impresión: Editorial Don Bosco-Centro Gráfico Salesiano - Telf.: 2831745
Tiraje: 300
Impreso en Ecuador - *Printed in Ecuador*

2015

Este libro ha sido debidamente examinado y valorado por evaluadores ajenos a PYDLOS EDICIONES, con el fin de garantizar la calidad científica del mismo.

El presente texto ha servido de referencia durante el Curso de formación de Posgrado: "Investigación Aplicada en Ciencias Sociales: Técnicas de producción de datos y análisis", actividad académica organizada por el Grupo de investigación PYDLOS delDepartamento de Investigación "Espacio y Población", en coordinación con las Facultades de CienciasEconómicas y Administrativas, Jurisprudencia, Psicología, Filosofía Letras y Ciencias de la Educación, y con aval de la DIUC de la Universidad de Cuenca.

ÍNDICE

PRESENTACIÓN

Durante años, el Programa Interdisciplinario de Población y Desarrollo Local Sustentable (PYDLOS) de la Universidad de Cuenca (Ecuador), y el grupo de OBETS, del Instituto de Desarrollo Social y Paz de la Universidad de Alicante (España) vienen manteniendo una estrecha colaboración en docencia e investigación.

Esta alianza ha sido posible gracias a la intensa actividad desarrollada por el director de PYDLOS, el profesor Dr. Alejandro Guillén, gracias a cuyo apoyo se han plasmado diversas líneas de cooperación.

En este marco, el presente texto es una prueba más ello, formando parte del conjunto de libros metodológicos dentro del compromiso entre la Universidad de Alicante y la Universidad de Cuenca, en materia de formación de Postgrado, a través del Grupo de Investigación PYDLOS del Departamento de Investigación "Espacio y Población", en coordinación con las Facultades de Ciencias Económicas y Administrativas, Jurisprudencia, Psicología, Filosofía Letras y Ciencias de la Educación, y con aval de la DIUC de la Universidad de Cuenca.

Estamos convencidos de que estos textos constituyen un aporte significativo para la formación de investigadores y docentes de nuestras universidades y para el robustecimiento de los vínculos entre ambas.

LOS AUTORES
Sant Vicent del Raspeig (Alicante)
Octubre 2015

1. LA INVESTIGACIÓN SOCIAL Y LA MEDICIÓN

La investigación científica de la realidad social siempre opera con algún nivel de medición. Desde el más básico, empleando la clasificación, hasta los más sofisticados que intentan establecer algún tipo de magnitud asociada a los fenómenos sociales. Una tarea tan simple como pueda ser crear categorías (poner un nombre a algo, y emplearlo para identificar y diferenciar) es un acto de medición. La sociedad misma ejerce esa capacidad de forma espontánea: crear tipos o categorías. Son recursos sociales para poder dar orientaciones y criterios de comportamiento a los individuos. Por lo general, esas tipologías sociales vienen caracterizadas por diferentes rasgos materiales (como posesiones), posiciones sociales (por ejemplo, según su posición en la cadena productiva), e incluso jerarquías de valores (que se les atribuyen culturalmente). Medir, en su función más básica de diferenciación, categorización, comparación y clasificación, es un procedimiento esencial y cotidiano en las sociedades. ¿Qué es más igual? ¿Qué es más diferente? Son operaciones básicas de la vida cotidiana.

Desde el punto de vista científico, la medición se efectúa construyendo sistemas de indicadores e índices que representen aspectos sustantivos de la realidad social. En esta labor, la estadística ejerce un papel instrumental importante. En el caso de la construcción de índices, la estadística multivariante facilita dos operaciones muy importantes. Primero, al ofrecer un sistema para determinar un índice que resuma la información que pueda contenerse en un conjunto de indicadores. En segundo lugar, al permitir la medición y establecimiento de índices a partir de realidades subjetivas que son difíciles observar de forma objetiva.

Cuando los rasgos que sirven de base para categorizar la realidad son externos (color de la piel, forma de vestir, poseer una vivienda, etc.), es relativamente fácil establecer un sistema de reglas para establecer indicadores, medir y calcular índices. Sin embargo, cuando la realidad que se desea medir no se percibe por los sentidos, es necesario establecer procedimientos diferentes. Ya no basta con criterios simples de indicadores directos. En estos casos, la

categorización (es decir, decidir cuáles son las categorías realmente existentes y significativas) pasan a primer plano. Medir variables e índices que expresen estados subjetivos de los individuos presenta sus propios desafíos. Este tipo de variables son de carácter subjetivo, al igual que lo son sus unidades de medida o los valores que adoptan esas unidades de medida. Existen en la mente de los individuos, pero no tienen una existencia objetiva, perceptible directamente por los sentidos. Un ejemplo, en otro ámbito de conocimiento, es la temperatura corporal. La percepción que cada individuo tiene de su temperatura corporal es subjetiva y personal. Para establecer una medición común y estándar para todos los individuos, debe construirse un instrumento externo como es el termómetro, que permite expresar de forma objetiva y comparable (mediante un instrumento o aparataje) esa temperatura corporal. La existencia de un termómetro no elimina la experiencia personal de cada individuo, las sensaciones subjetivas que le pueden producir la sensación de calor o frío. Lo que permite es establecer un indicador que ofrece una información externa y objetiva de la temperatura corporal. Y a partir de ello, facilitar la comparación intersubjetiva. En ambas situaciones, dónde la medición es de características externas (observables directamente) o se refiere a características internas (no observables directamente), las respuestas a cómo construir índices de medición deben ser diferentes, si bien respetando principios metodológicos semejantes.

Cuando hablamos de medir, en cualquiera de sus niveles, resulta evidente que existe un lenguaje apropiado, que no es el lenguaje natural. Las operaciones que se efectúan sobre las mediciones, ya sea con la finalidad de descripción o explicación, requieren de un lenguaje formalizado creado para ello: el lenguaje matemático y estadístico. La estadística y las matemáticas en general, son el lenguaje que opera con mediciones cuantitativas. Desde la más básica de clasificar, hasta las explicaciones empíricas más sofisticadas.

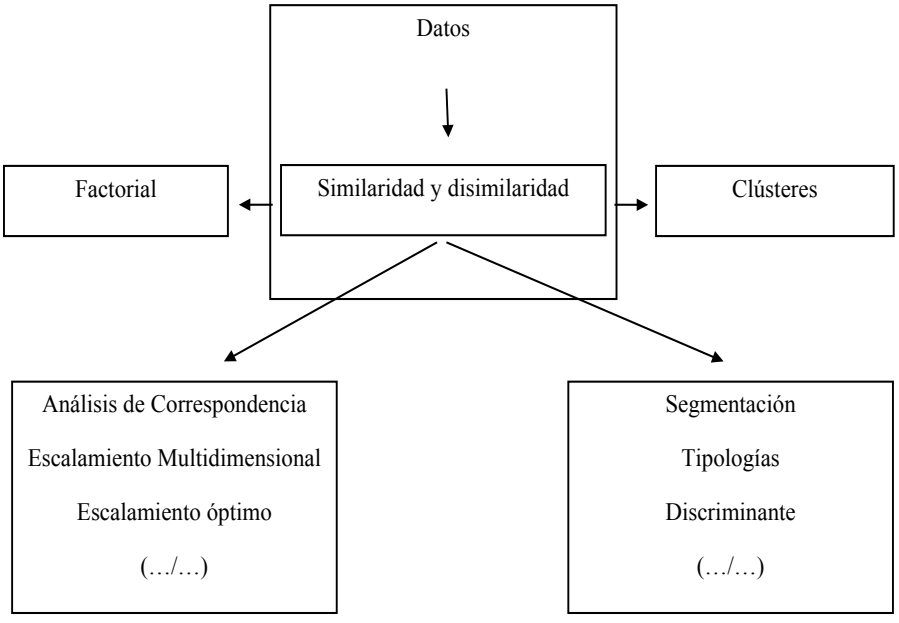
En este texto presentamos tres estrategias de medición multivariante, que son útiles tanto para mediciones de características que se pueden percibir direc-tamente, como de estados subjetivos. Nos centraremos en su empleo para medir (por ejemplo, estados subjetivos de los individuos), permitiendo construir, por ejemplo, un índice. Resulta evidente, una vez que hemos logrado asignar una cifra a cada individuo o caso (su valor en un índice), ya es decisión del investigador si las empleará para agrupar los casos en tipologías o usar su magnitud para relacionarla con otras magnitudes medidas en otras variables.

Es importante que el investigador comprenda que la mayoría de los procedimientos estadísticos están interconectados entre sí. Podemos llamarlos de forma diferente por motivos varios. Así es habitual que sea la finalidad inicial para la que se establecieron la que los diferencia y les da nombre. Sin embar-

go, con un análisis factorial (por ejemplo), una vez efectuado, se pueden atribuir magnitudes a los sujetos. A partir de ellas se les puede clasificar o efectuar otras muchas operaciones. Una vez que hemos "medido" los sujetos respecto a algo, las demás operaciones son derivadas de esa labor esencial. Por eso, existen muchas vías alternativas (procedimientos estadísticos) que pueden dar respuesta a una misma pregunta. Como hemos dicho, en esta ocasión vamos a considerar la medición tanto desde el punto de vista de los estados subjetivos como objetivos. Al fin y al cabo, las valoraciones que efectúan los individuos toman como referencia sus escalas personales de carácter subjetivo.

En este libro vamos a partir de los procedimientos para determinar distancias o proximidades (similitud y disimilitud) entre casos o variables. Tras ello, presentaremos en términos de distancias y proximidades (especialmente entre casos, aunque no solamente), el análisis de conglomerados (clústeres), así como la utilidad que tiene el análisis de varianza en la definición del número de clústeres. En lo que se refiere a la similitud o disimilitud entre variables, mostramos el procedimiento estadístico denominado análisis factorial, junto a los procedimientos para determinar la fiabilidad de la medición (alfa de Cronbach).

Cuadro 1. Ejemplos de modelos basados en la determinación de distancias



Dos de los procedimientos multivariantes principales en la construcción de índices son, por un lado el análisis factorial, y por otro el escalamiento multidimensional. Tanto el análisis factorial como el escalamiento multidimensional tienen su origen en la psicometría. El análisis factorial, para efectuar mediciones empleando múltiples ítems (incrementando así la fiabilidad). El escalamiento multidimensional, por otra parte, fue desarrollado para ayudar a comprender las valoraciones que un conjunto de individuos efectuaban sobre la similitud o disimilitud entre un grupo de objetos. En el primer caso se utilizan datos multivariantes para, mediante el coeficiente de correlación, estimar las proximidades (similaridad) entre variables. En el segundo se utilizan medidas de distancia entre objetos, basadas en la similaridad o la disimilaridad que se aprecie entre ellos. El escalamiento multidimensional utiliza los datos que expresan similaridad o disimilaridad entre los objetos como parte del procedimiento para construir escalas objetivas asociadas a atributos subjetivos.

Muchos procedimientos estadísticos, tales como el análisis factorial, el análisis de conglomerados, o el escalamiento multidimensional tienen como punto de partida la matriz que define las distancias[1] o las proximidades entre pares de casos o variables. A partir de dichas matrices de distancias, se procede a formar los conglomerados, se extraen los factores o se identifican las estructuras y dimensiones presentes en los datos.

1. Las distancias son el punto de partida en el análisis de Conglomerados (las distancias entre casos o variables), en el escalamiento multidimensional (distancias entre casos o también entre variables), en el análisis factorial (la similitud entre variables define el factor). En los tres métodos, la similitud o la disimilitud son muy importantes dado que los casos son agrupados en función a su proximidad.

2. LA SIMILITUD Y LA DIFERENCIA

Los conceptos de similitud y disimilitud son esenciales en la investigación social. El nivel más básico de medición, el nominal, emplea la clasificación de objetos (cualidades o características de un objeto). El criterio de clasificación es la similitud o la disimilitud. En este contexto, los conceptos de similitud y proximidad se consideran sinónimos, al igual que el de distancia y disimilaridad. Ambos conceptos expresan una misma realidad desde dos puntos de vista opuestos. Mientras que el concepto de distancia expresa en qué medida son diferentes dos objetos, el concepto de similaridad mide el grado de proximidad entre ellos. En ese sentido, para dos casos que sean muy parecidos, la medida de distancia debería de ser pequeña mientras que, por el contrario la medida de similaridad debería de ser elevada. La idea de medir la similitud o disimilitud (la proximidad o distancia) entre objetos (casos) o variables es el punto de partida de muchas técnicas estadísticas.

Vamos a plantear un ejemplo sencillo de cómo se calcula una distancia. Se pueden emplear medidas muy diferentes para cuantificar la similaridad o la disimilaridad. Una de las más utilizadas para medir la distancia entre *casos* (objetos) es la distancia euclidiana al cuadrado. En el caso de *variables*, el coeficiente de correlación es uno de los que se utiliza con más frecuencia para medir la similaridad entre dos variables. Cuando el valor de correlación es muy elevado indica que las dos variables son muy parecidas.

Vamos a considerar uno de los índices de Desarrollo Democrático en Latinoamérica. Estimaremos qué distancia existe entre dos casos (en este ejemplo países). Para ello, operaremos con los valores que poseen en cuatro variables diferentes (realmente son dimensiones que sintetizan diferentes indicadores). La idea es combinar la información que facilitan las cuatro dimensiones para determinar en qué forma ambos casos (dos países) se parecen o son diferentes. Tomemos como ejemplo de índice de disimilitud la distancia euclídea al cuadrado. Este índice en definitiva lo que hace calcular las diferencias que existen entre los valores que tienen los casos en las variables consideradas, elevarlos al cuadrado y sumar los términos.

Tabla 1. Las cuatro dimensiones[2] del IDD-lat 2014

País	DIM I	DIM II	DIM III	DIM IV
Uruguay	8	9,5	0,7	1,3
Costa Rica	7,43	6,7	2,3	0,1
Chile	7,2	5,3	2,4	1,7
Argentina	6,5	3,1	2,6	0,7
Bolivia	6	2,3	-0,25	-1,5
Perú	5,4	4	1,5	1,6
Nicaragua	5,3	3	-2,3	-0,8
Ecuador	5,3	2,9	0,4	0,3
Brasil	5,1	2,4	0,9	-0,3
El Salvador	5	6,9	-0,7	-1,8
Paraguay	5	3,3	-1,3	-0,7
Panamá	4,8	2,5	1,4	0,4
Rep. Dominicana	4,7	2,8	-3,1	-0,9
México	4	3,7	0,4	1,6
Venezuela	3,1	1,9	-0,7	0,3
Colombia	2,8	3,5	-1	0,9
Honduras	2,4	3,6	-1	-1,2
Guatemala	1	3,8	-2,3	-1,7

Fuente: Datos del IDD-Lat 2014

2. Dimensión I: "Democracia de los ciudadanos". Evalúa el respeto de los derechos políticos y las libertades civiles. Dimensión II: "Democracia de las instituciones". Mide la calidad institucional y la eficiencia del sistema político. Dimensión III. "Democracia social y humana". Analiza la capacidad del sistema democrático para generar políticas que aseguren bienestar y desarrollo humano. Dimensión IV. "Democracia económica".

Para determinar la distancia euclídea al cuadrado entre Ecuador y Bolivia, por ejemplo, procederíamos de la forma siguiente.

País	DIM I	DIM II	DIM III	DIM IV
Bolivia	6	2,3	-0,25	-1,5
Ecuador	5,3	2,9	0,4	0,3

la distancia euclídea al cuadrado es simplemente la suma de las diferencias al cuadrado.

Distancia euclídea al cuadrado =

$(6 - 5,3)^2 + (2,3 - 2,9)^2 + (-0,25 - 0,4)^2 + (-1,5 - 0,3)^2 =$

$(0,7)^2 + (-0,6)^2 + (-0,65)^2 + (-1,8)^2 =$

$(0,49) + (0,36) + (0,42) + (3,24) = 4,51$

Es decir, de acuerdo a esta medida de distancia, la disimilitud entre Bolivia y Ecuador en Desarrollo Democrático es de 4,51. Si reiteramos este procedimiento para todos los pares que se pueden formar entre países, es factible calcular una matriz con la distancia entre todos los países. Esta matriz de distancias es la base para efectuar análisis multivariantes con diferentes intencionalidades. Es evidente que la elección de una medida de disimilaridad u otra debe hacerse con una justificación teórica. No debemos olvidar que el emplear una medida u otra puede tener consecuencias sobre los resultados. En las páginas siguientes vamos a considerar qué elementos deben de considerarse para la elección de una mediada de distancia.

Existen al menos tres decisiones que se deben de tomar cuando se decide trabajar con índices de similaridad o disimilaridad. Nos detendremos con más detalle en ampliar el segundo y el tercero, donde, por lo general, cabe una intervención mayor del investigador.

a) La primera decisión se refiere al tipo de datos que estamos utilizando. El nivel de medición de los datos sugerirá qué tipo de distancia debe tomarse en consideración, y con ello el índice que puede ser el más adecuado. Así consideraremos tres tipos diferentes de datos:

- los datos expresan mediciones con nivel de intervalo,
- los datos expresan frecuencias,
- y los datos son de carácter binario (expresando la presencia o ausencia de una característica o cualidad).

b) La segunda decisión importante se refiere a la conveniencia o no, de normalizar (o estandarizar) los datos. Es decir, los valores que adoptan los casos. Por lo general, la estandarización y la normalización se aplican a los datos expresados con nivel de medición de intervalo, o como frecuencias. *La estandarización establece un valor para la media y la varianza de la variable (es decir, modifica la distribución) mientras que la normalización modifica los valores para re-expresarlos dentro de un nuevo rango de variabilidad*. Debemos recordar que la diferente unidad de medida en que se expresen los datos (euros, años, etc.) hace que las variables con magnitudes que pueden alcanzar valores elevados contribuyan en mayor grado al índice de disimilaridad.

Un procedimiento para evitar, o al menos atenuar, el impacto que las diferentes unidades de medición de las variables pueden tener en la estimación de la distancia o similitud entre casos, es expresar todas las variables en la misma unidad. Éste procedimiento se denomina normalización, y existen diferentes procedimientos para lograr este objetivo. La normalización se efectúa mediante transformaciones y permite hacer comparables los valores de los datos antes de calcular proximidades. La normalización y la estandarización puede realizarse, a) para todos los valores que adopte un caso en un conjunto de variables (es decir, normalizamos los valores de un caso tomando como referencia los valores de ese mismo caso en las distintas variables), o b) normalizando la variable tomando como referencia los valores de todos los casos en esa variable.

a) Estandarización de un caso mediante puntuación Z tomando como referencia un conjunto de seis variables:

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6
Caso 1	15	2	33	105	22	11

(Vector fila)

Media: 31,3

Desviación típica: 37,5

Puntuación Z = (x-media)/desviación típica

Z de la variable 1 para el caso 1 = (15 – 31,3) / 37,5 = -,43

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6
Caso 1	15	2	33	105	22	11
Caso1[(a)] "Z"	-,43	-,78	,04	1,9	-,24	-,54

a. Expresado con puntuación Z

b) Estandarización de un caso mediante puntuación Z tomando como referencia los valores de los siete casos en una variable:

Casos	Variable 1
Caso 1	13
Caso 2	25
Caso 3	32
Caso 4	12
Caso 5	56
Caso 6	43
Caso 7	15

(Vector columna)

Media: 28

Desviación típica: 16,7

Puntuación Z del caso 1 en la variable 1 = (13 – 28) / 16,7 = -,89

Casos	Variable 1	Variable 1[(a)] "Z"
Caso 1	13	-,89536
Caso 2	25	-,17907
Caso 3	32	,23876
Caso 4	12	-,95505
Caso 5	56	1,67133
Caso 6	43	,89536
Caso 7	15	-,77598

a. Expresado con puntuación Z

Entre las transformaciones de estandarización más habituales se encuentran las puntuaciones Z, que ya hemos visto, así como tomar de referencia una media de 1, o una desviación típica de 1 (a diferencia de Z, que fija una media de 0 y una desviación típica de 1). Por ejemplo, para estandarizar las distribuciones de forma que tengan una media de 1, se dividen todos los valores por la media. De forma extensiva, de desear estandarizar las distribuciones para que todas tengan una desviación típica de 1, se dividen todos los

valores de la variable por la desviación típica. Es importante tener presente, en este caso, que se producen transformaciones parciales de las distribuciones.

Procedimientos usuales de estandarización y normalización (casos y variables)	
Estandarización (afecta a la distribución)	
Media de 0 y desviación típica de 1	Restar cada valor a la media y dividir por la desviación típica
Media de 1	Dividir todos los valores por la media
Desviación típica de 1	Dividir todos los valores por la desviación típica
Normalización (no afecta a la distribución)	
Rango de -1 a 1	Dividir todos los valores por el rango de la variable
Rango de 0 a 1	Restar a todos los valores el valor más pequeño y dividir por el rango de la variable
Valor máximo de 1	Dividir todos los valores por el valor mayor de la distribución

Otra opción es *normalizar* las variables, por ejemplo fijando un rango de variabilidad entre –1 a 1, un rango 0 a 1, o establecer como magnitud máxima la de 1. *En este caso, debemos recordar que las medias o las desviaciones típicas de las variables continúan siendo diferentes*. La normalización de la variabilidad de una variable puede limitar sus valores entre -1 y 1. Para ello, todos los valores que van a ser normalizados se dividen por el rango de la variable. Otra opción es establecer una variabilidad normalizada para todas las variables entre 0 y 1. Así, en el ejemplo anterior, normalizamos entre 0 y 1 de la forma siguiente. A cada caso se les resta el valor menor de la distribución y se divide por el rango.

El rango varía entre 12 (el menor valor) y 56 (el mayor valor). Así 56 – 12 = 44. A cada valor se le resta el menor. Por ejemplo, el caso 4 tiene el menor valor, 12. Si a 12 se le resta 12= 0. Como es conocido 0/44 = 0. En el otro extremo, tomemos el valor mayor, que en este ejemplo es el caso 5. El valor del caso 5 es 56, le restamos el menor que es 12 = 44. Y si dividimos

Casos	Variable 1
Caso 1	13
Caso 2	25
Caso 3	32
Caso 4	12
Caso 5	56
Caso 6	43
Caso 7	15

44 entre el rango (44), tenemos 44/44 = 1. Cualquier otro valor quedará transformado entre dichos extremos, y con ello normalizado en su cuantía. Es decir, comparable con otras variables normalizadas de igual modo. El caso 2, con un valor de 25, se normaliza en (25-12)/44 = 0,29.

Para establecer que el valor máximo de la distribución sea 1 se debe dividir todos los valores en la variable original por el valor más elevado que contenga (valor máximo). Todos estos procedimientos de estandarización y normalización pueden efectuarse tomando como referencia tanto la variable, como los valores de un caso, en la forma que hemos considerado previamente.

La ventaja evidente de la normalización y la estandarización es que facilita la comparación, de modo que los índices de disimilitud o similitud se ven menos influenciados por las unidades de medida de cada variable. Por ello, lo habitual es efectuar las transformaciones previamente al cálculo de los índices de disimilitud o de proximidad.

Sin embargo, es importante considerar que los efectos de los casos extremos se intensifican con las transformaciones de normalización. De esta forma, los casos extremos provocan que se comprima el rango de variabilidad de los valores que pueden considerarse más usuales en esa medición. Es una diferencia significativa entre normalización y estandarización: la normalización establece límites de variación (-1 a 1, de 0 a 1, etcétera) que reduce su variabilidad cuando se presentan casos extremos.

No siempre puede resultar interesante el normalizar o estandarizar las variables. Lo importante es atender al significado de las variables, dado que determinadas unidades de medida, y la variabilidad que implica, puede expresar elementos sustantivos en lo que se refiere a la naturaleza de los fenómenos que expresan dichas variables. Es decir, pueden darse ocasiones en que deseamos que determinadas variables tengan un peso especial en la determinación de la similitud o la disimilitud. En esa situación, donde interesa con-

servar un peso especial en algunas variables, no es conveniente la normalización. Introduciremos ejemplos más adelante.

c) Por último, la tercera decisión se refiere a la necesidad que exista, o no, de *normalizar el índice* que expresa de forma resumida la diferencia o similitud entre los casos (o las variables). La intención es que los valores que adopte el índice sean más comprensibles para el investigador. Estas transformaciones se efectúan sobre el índice que calcula la distancia final, es decir, que la normalización se aplica después de calcular la medida de distancia. Algunas opciones habituales son: Valores absolutos, Cambiar el signo, y Cambiar la escala al rango de 0 hasta 1.

En el caso de tomar el *valor absoluto* del índice de disimilaridad o similaridad, estaremos indicando que el posible signo que adopte el coeficiente no tiene significado relevante para el investigador. Es el caso del coeficiente de correlación, usado para expresar proximidad. En el caso que solamente interese su magnitud como referente de proximidad, el signo es perfectamente prescindible. Recordemos que en este caso, el coeficiente de correlación se encuentra normalizado entre -1 y +1. Tomar su valor absoluto lo transforma en un nuevo rango, entre 0 y 1. La opción de *cambiar de signo*, esencialmente es un cambio de tipo de medida. Transforma las mediciones de disimilaridad en similaridad. Su consecuencia es que invierte el orden de las distancias entre los casos o las variables. Por último, los índices de disimilaridad o similaridad pueden normalizarse entre 0 y 1. Para ello, tomados los valores de las mediciones de similaridad (o disimilaridad), se resta de todas ellas el valor de la distancia menor, y se dividen por el rango de variabilidad (es decir, el valor de la distancia mayor entre dos casos o variables, menos el valor de la distancia menor entre dos casos o variables). Mediante este procedimiento, las distancias calculadas entre casos o variables se normalizan a una variación entre 0 y 1.

Es importante notar la diferencia entre las dos transformaciones mencionadas. La primera, (apartado b) se refiere a las transformaciones de los valores que presentan los casos, de forma que la unidad de medida tenga un impacto menor en los coeficientes de disimilaridad o similaridad. La segunda (apartado c), modifica los resultados de la aplicación de cualquiera de los índices de proximidad o distancia. Es decir, una matriz de distancias se vería transformada de forma que todas las distancias oscilarán entre 0 y 1.

2.1. Los diferentes tipos de índices y su medición

Como es evidente existen muchas definiciones operacionales diferentes, en términos matemáticos y estadísticos, para medir los conceptos de distancia

y similaridad. Vamos seguidamente a presentar diferentes procedimientos para calcular coeficientes alternativos en la medición de distancias o proximidades. Para ello organizaremos la presentación según 1) el nivel de medición en que están expresados los datos, y 2) si está diseñado para medir similaridad o disimilaridad (proximidad o distancia).

2.1.1. *Datos expresando nivel de medición de intervalo*

a) Disimilaridad

Cuando los datos están expresados en un nivel de medición de intervalo o superior, y consideramos la estimación de índices basados en la *disimilitud, distancia o diferencia* pueden considerarse las siguientes opciones de medición.

Distancia euclídea

Esta medición fue presentada como ejemplo en líneas anteriores. La distancia entre dos objetos, X e Y, es la raíz cuadrada de la suma de las diferencias al cuadrado de los valores.

$$\text{Distancia euclídea } (x,y) = \sqrt{\Sigma_i (x_i - y_i)^2}$$

Para determinar la distancia euclídea entre Ecuador y Bolivia, por ejemplo, procederíamos de la forma siguiente.

País	DIM I	DIM II	DIM III	DIM IV
Bolivia	6	2,3	-0,25	-1,5
Ecuador	5,3	2,9	0,4	0,3

la distancia euclidea es simplemente la raíz cuadrada de suma de las diferencias al cuadrado.

Distancia euclídea =

$$\sqrt{4,51} = 2,12$$

La distancia euclídea, presenta dos inconvenientes. En primer lugar, la distancia depende de las unidades que se empleen para expresar las variables o indicadores. Esto significa que los cambios de escala en las variables implican cambios en las distancias. Una forma de evitar este efecto es transformar y normalizar las variables. Otra consecuencia es su sensibilidad a la posibilidad de que las variables ofrezcan medidas redundantes (es decir, que estén altamente correlacionadas). En estas situaciones, la distancia euclídea sobreestima la disimilaridad o distancia entre los individuos. Una posible solución a esto es extraer componentes principales de las variables o indicadores (que

como sabemos estarían incorrelacionados) y utilizarlos para la estimación de disimilaridad, en lugar de las variables o indicadores originales. Otra alternativa es ponderar, con pesos inversamente proporcionales a la correlación entre dos variables, la distancia estimada para cada par de ellas.

Todos estos comentarios sugieren que esta medida de distancia es recomendable cuando las variables estén medidas en unidades semejantes. Es decir, cuando la escala es homogénea. La medición de la distancia euclídea, cuando se aplica a la distancia entre varias dimensiones (o variables), define una distancia geométrica dentro de ese espacio multidimensional. *Por lo general se ve poco influenciada por la presencia de casos extremos, siendo muy sensible, como sabemos, a las diferencias en escala.*

Distancia euclídea al cuadrado

Recordemos que consiste simplemente en la suma de las diferencias al cuadrado de los valores.

$$\text{Distancia euclídea cuadrada } (x,y) = \Sigma_i (x_i - y_i)^2$$

La distancia euclídea al cuadrado tiene como desventaja que también depende de la unidad de medición, es decir la magnitud en que se exprese la variable. Por ejemplo si la variable ingresos se encuentra medida en euros o dólares, la diferencia en esa variable para dos casos, será siempre superior en magnitud, a la diferencia que pudiese darse al comparar la variable edad. En ese sentido, la magnitud con que se expresa la medición, es decir los valores que puede llegar a alcanzar, influye en la determinación de la similitud o distancia entre diferentes casos para estimar determinadas "distancias". *La medición concede progresivamente un peso cada vez mayor a los objetos, cuanto más separados están. En ese sentido, potencia las diferencias.*

Bloque, Manhattan

Otra forma de medir las distancias entre dos objetos es utilizar los valores absolutos, que resultan de restar los valores de un caso (en una variable) con los valores de otro caso (en esa misma variable), en lugar de emplear sus cuadrados. Es el caso de la denominada Bloque, (también llamada distancia Manhattan o Ciudad). Es simplemente la suma de las diferencias absolutas de los valores, en las variables consideradas, para cada par de casos (objetos). En la medida que las diferencias no se han elevado al cuadrado, las diferencias más importantes no tendrán tanto peso como sucede cuando se elevan al cuadrado.

$$\text{Distancia } (x,y) = \Sigma_i \left| x_i - y_i \right|$$

País	DIM I	DIM II	DIM III	DIM IV
Bolivia	6	2,3	-0,25	-1,5
Ecuador	5,3	2,9	0,4	0,3

Distancia Bloque =

$$|6 - 5{,}3| + |2{,}3 - 2{,}9| + |0{,}25 - 0{,}4| + |-1{,}5 - 0{,}3| =$$

$$0{,}7 + 0{,}6 + 0{,}65 + 1{,}8 = 3{,}75$$

La distancia entre Bolivia y Ecuador en el índice de democracia sería de 3,7 de utilizar esta medición. En conjunto, la distancia viene a definir la distancia media entre dimensiones, con unos *resultados bastante próximos a los de la distancia euclídea, si bien el efecto de los casos extremos es menos fuerte dado que no se elevan los valores al cuadrado*.

Chebychev

La distancia Chebychev también emplea las diferencias expresadas en valores absolutos. Sin embargo, no utiliza todas las variables. Esta distancia se define como la diferencia mayor en términos absolutos entre dos casos, considerando todas las diferencias entre variables. En ese sentido, *ignora gran parte de la información disponible*. *Solamente la variable que muestra la diferencia máxima entre los dos casos será la utilizada*.

$$\text{Distancia } (x,y) = \text{MA } X_i \ |x_i - y_i|$$

País	DIM I	DIM II	DIM III	DIM IV
Bolivia	6	2,3	-0,25	-1,5
Ecuador	5,3	2,9	0,4	0,3

Distancia Chebychev =

$$|6 - 5{,}3| \ , \ |2{,}3 - 2{,}9| \ , \ |0{,}25 - 0{,}4| \ , \ |-1{,}5 - 0{,}3| =$$

$$0{,}7 \ , 0{,}6 \ , 0{,}65 \ , 1{,}8 = 1{,}8$$

Aquellas observaciones que hacíamos anteriormente respecto al peso en el índice que pueden llegar a tener las variables con una unidad de medida superior (por ejemplo renta cuando la utilizamos a la par que la variable edad) son especialmente relevantes en esta ocasión. Dado que se utiliza solamente una variable como referencia de distancia, será aquella que emplea magnitudes mayores la que defina finalmente el valor del índice. Supongamos que empleamos las variables Producto Interior Bruto (en miles de millones),

Porcentaje de participación en las elecciones generales y número de desempleados. La distancia que se calcule empleando este índice tendrá en cuenta exclusivamente el PIB, dado que arrojará la diferencia de mayor magnitud. Por ello, es conveniente efectuar previamente transformaciones en los datos, o tener una razón significativa para utilizar este índice. *La lógica que rige esta medición de distancia es que lo importante es la diferencia, sin dar importancia a las dimensiones o variables que definen esas diferencias*. En ese sentido, la diferencia es lo central, concediendo un peso o importancia equivalente a todas las dimensiones.

Son muchas las alternativas de medición de disimilitud o distancia, y cada una de ellas responde a una lógica diferente. Dado el carácter introductorio de este texto, donde se muestra la lógica de la estimación de distancias, no abundamos en más ejemplos de disimilaridad en mediciones métricas. Como puede apreciarse en la tabla comparativa siguiente, los diferentes métodos de estimar las distancias ofrecen magnitudes diferentes. Especialmente el resultado de la distancia euclídea al cuadrado. Esto es cierto tanto en términos absolutos (puede afectar a la representación de los datos) como en términos relativos, de distancia entre ellos. Elevar al cuadrado incrementa la separación entre casos. No obstante, la posición ordinal de los países con respecto a Ecuador (tomado como referencia) no se ve alterada. Sin embargo, con el empleo de la medida "bloque o Manhattan" se puede afectar tanto a la posición ordinal de algún país (así como a su distancia) con Ecuador. Así, Nicaragua (3,9) estaría a menor distancia de Ecuador que Argentina (4) o Venezuela (4,3) si empleamos la distancia "bloque".

Tabla 2. Cálculo de la disimilaridad de Ecuador con otros países según diferentes coeficientes.
Distancia de Ecuador con varios países en el IDD-lat 2014

	Distancia euclídea	Distancia euclídea al cuadrado	Distancia de Chebychev	Distancia de bloques de ciudad
Brasil	0,949	0,9	0,6	1,8
Panamá	1,192	1,42	1	2
México	2,005	4,02	1,3	3,4
Perú	2,03	4,12	1,3	3,6
Paraguay	2,035	4,14	1,7	3,4
Bolivia	2,124	4,513	1,8	3,75
Argentina	2,546	6,48	2,2	4

	Distancia euclídea	Distancia euclídea al cuadrado	Distancia de Chebychev	Distancia de bloques de ciudad
Venezuela	2,655	7,05	2,2	4,3
Nicaragua	2,917	8,51	2,7	3,9
Colombia	2,988	8,93	2,5	5,1
Honduras	3,621	13,11	2,9	6,5
Rep. Dominicana	3,75	14,06	3,5	5,4
Chile	3,915	15,33	2,4	7,7
El Salvador	4,659	21,71	4	7,5
Costa Rica	4,757	22,627	3,8	8,03
Guatemala	5,531	30,59	4,3	9,9
Uruguay	7,207	51,94	6,6	10,6

Fuente: elaboración propia

Hasta aquí hemos operado calculando las distancias entre casos (países). El mismo procedimiento puede aplicarse a las variables. Es decir, podemos aplicar todos estos índices de distancia para determinar la similitud entre las variables.

Tabla 3. Las cuatro dimensiones[3] del IDD-lat 2014

País	DIM I	DIM II	DIM III	DIM IV
Uruguay	8	9,5	0,7	1,3
Costa Rica	7,43	6,7	2,3	0,1
Chile	7,2	5,3	2,4	1,7
Argentina	6,5	3,1	2,6	0,7
Bolivia	6	2,3	-0,25	-1,5
Perú	5,4	4	1,5	1,6

3. Dimensión I: "Democracia de los ciudadanos". Evalúa el respeto de los derechos políticos y las libertades civiles. Dimensión II: "Democracia de las instituciones". Mide la calidad institucional y la eficiencia del sistema político. Dimensión III. "Democracia social y humana". Analiza la capacidad del sistema democrático para generar políticas que aseguren bienestar y desarrollo humano. Dimensión IV. "Democracia económica".

País	DIM I	DIM II	DIM III	DIM IV
Nicaragua	5,3	3	-2,3	-0,8
Ecuador	5,3	2,9	0,4	0,3
Brasil	5,1	2,4	0,9	-0,3
El Salvador	5	6,9	-0,7	-1,8
Paraguay	5	3,3	-1,3	-0,7
Panamá	4,8	2,5	1,4	0,4
Rep. Dominicana	4,7	2,8	-3,1	-0,9
México	4	3,7	0,4	1,6
Venezuela	3,1	1,9	-0,7	0,3
Colombia	2,8	3,5	-1	0,9
Honduras	2,4	3,6	-1	-1,2
Guatemala	1	3,8	-2,3	-1,7

Fuente: Datos del IDD-Lat 2014

Así, como ejemplo, para calcular la disimilitud entre la DIM 1 y la DIM 2, mediante distancia euclídea al cuadrado, calcularíamos las diferencias entre pares de valores de un caso en las dos variables.

Distancia Euclidea al cuadrado entre DIM 1 y DIM 2:

$$(8 - 9{,}5)^2 + (7{,}43 - 6{,}7)^2 + (7{,}2 - 5{,}3)^2 + \ldots\ldots + (2{,}4 - 3{,}6)^2 + (1 - 3{,}8)^2 = 78{,}6$$

Distancia Euclidea entre DIM 1 y DIM 2 = 8,86

Distancia Bloque entre DIM 1 y DIM 2 = 34

Distancia Chebychev entre DIM 1 y DIM 2 = 3,7

Nuevamente, si calculamos las distancias entre las cuatro variables (en este caso dimensiones) tendremos una matriz de distancias entre las cuatro variables. La tabla 4 muestra la distancia euclídea entre las cuatro variables.

Como hemos podido apreciar, la distancia se calcula a partir de los vectores de valores asociados a los casos (vectores fila en una matriz rectangular), o a las variables en columnas (vectores columna en la matriz rectangular). Las diferentes formas de estimar las distancias son factibles de aplicarse para comparar casos o variables.

Tabla 4. Matriz de distancias

	Distancia euclídea			
	DIM 1	DIM 2	DIM 3	DIM 4
DIM 1	,000			
DIM 2	8,868	,000		
DIM 3	21,918	19,131	,000	
DIM 4	22,169	18,842	5,332	,000

Esta es una matriz de disimilaridades

Vamos a recordar las transformaciones para normalizar distancias. Así, si queremos normalizar la matriz anterior en un rango entre 0 y 1, le restaremos a todas las distancias el valor de la distancia menor, y posteriormente las dividiremos por el rango. La distancia menor es 5,3. El rango es el resultado de restar a 22,1 (distancia mayor) la distancia 5,3 (distancia menor), lo que es igual a 16,8. Así, normalizar la distancia entre la DIM 1 y la DIM 2 implica tomar el valor 8,8 y restarle 5,3 = 3,5. Después dividir por el rango: 3,5/16,8 = 0,20. La matriz normalizada en el rango 0 hasta 1 es la siguiente.

Tabla 5 Matriz de distancias normalizada 0-1

	Distancia euclídea			
	DIM 1	DIM 2	DIM 3	DIM 4
DIM 1	,0			
DIM 2	0,20	,0		
DIM 3	0,98	0,82	,0	
DIM 4	1	0,80	0	,0

Esta es una matriz de disimilaridades normalizada 0-1

Recordemos que, en este caso, hemos normalizado las distancias. Esta matriz puede ser empleada posteriormente para múltiples análisis multivariantes, basados en matrices de proximidad o distancia.

b) Similaridad

Cuando los datos están expresados en un nivel de medición de intervalo o superior, y consideramos la estimación de índices basados en las medidas

de similitud, proximidad o semejanza, las opciones más usuales son el coeficiente de correlación de Pearson y el Coseno.

Correlación de Pearson

El coeficiente de correlación de Pearson puede ser considerado como una medición de similaridad entre variables o entre casos medidos en un nivel de intervalo o superior. La proximidad o igualdad entre dos variables vendrían definidas por la correlación entre los vectores definidos por los valores de cada una de ellas. Como es bien conocido, el coeficiente de correlación de Pearson oscila entre -1 y +1 (es un coeficiente normalizado), dónde 0 expresa la ausencia de correlación entre las variables. Cuanto más próximo esté el coeficiente a -1 o +1, más fuerte es la relación entre las variables[4]. En otras palabras, cuanto más elevada es la correlación, tanto en positivo (directamente proporcional) como en negativo (inversamente proporcional), más fuerte es la relación, y expresa que las dos variables están bastante próximas. Una relación no significativa o muy baja indicaría que las dos variables son muy diferentes.

$$\text{Correlación } (x,y) = \Sigma_i \, Zx_i Zy_i / N - 1$$

Téngase presente que en este caso, la proximidad o similaridad se establece tanto entre variables, como entre casos. Depende del vector (fila o columna) que se emplee para estimar la correlación. En el caso de los índices de democracia considerados en el ejemplo, la correlación (proximidad) entre dimensiones es la siguiente

	Correlación entre vectores de valores (Columna)			
	DIM 1	DIM 2	DIM 3	DIM 4
DIM 1	1,000			
DIM 2	,497	1,000		
DIM 3	,622	,265	1,000	
DIM 4	,386	,193	,645	1,000

Matriz de similaridades

Puede observarse como la dimensión 3 (“Democracia social y humana”) y la dimensión 4 (“Democracia económica”) son las más próximas, mientras

4. Considerando siempre que el coeficiente de correlación sea significativo. Es decir, que dicha relación existe en la realidad según exprese su significación estadística.

que las dimensiones 2 ("Democracia de las instituciones") y la dimensión 4 ("Democracia económica") son las más disimilares, en la medición del Desarrollo Democrático. En cierto modo, expresa en qué modo los índices están más o menos próximos en la medición común que aspiran a realizar. Este tipo de análisis es bastante semejante (con las diferencias correspondientes) con el análisis factorial.

Y, procediendo del mismo modo para los vectores fila con los valores de cada caso, la matriz de correlaciones expresaría la similitud entre casos (en este ejemplo, países).

	Uruguay	Costa Rica	Chile	Argentina	Bolivia	Perú	Nicaragua	Ecuador
Uruguay	1							
Costa Rica	0,92	1						
Chile	0,88	0,96	1					
Argentina	0,65	0,85	0,92	1				
Bolivia	0,79	0,92	0,98	0,97	1			
Perú	0,90	0,93	0,99	0,88	0,96	1		
Nicaragua	0,90	0,89	0,95	0,81	0,92	0,98	1	
Ecuador	0,84	0,91	0,98	0,92	0,98	0,99	0,97	1

Matriz de similaridades
Correlación entre vectores de valores (fila)

Tomando como referencia las cuatro dimensiones consideradas, los dos países (de los analizados) más similares (próximos) son Perú y Ecuador (0,99), o Perú y Chile (0,99). Los menos similares, en este caso, son Argentina y Uruguay (0,65). Como puede apreciarse, el coeficiente de correlación expresa similaridad sobre la base de la asociación.

Coseno

Esta es una medida de similaridad que utiliza los cosenos de los vectores definidos por los valores de cada variable (vector columna), o de cada caso en las diferentes variables (vector fila). Desarrolla el planteamiento de expresar los datos como vectores, si bien en esta ocasión empleados para calcular la similitud. La similitud entre dos vectores, evaluada por el coseno del ángulo, oscila entre los valores -1 y 1. El valor máximo de 1 resulta cuando el ángulo entre los dos vectores es cero. En definitiva, que ambos vectores apuntan hacia la

misma posición, siendo paralelos. Cualquier otro ángulo ofrecería un valor inferior a 1. Cuando los vectores son ortogonales su coseno es cero y en el caso de apuntar en sentido opuesto alcanzaría un valor máximo de -1. La medida es independiente (excepto el signo) de la longitud de los vectores.

	Coseno de vectores de valores			
	DIM 1	DIM 2	DIM 3	DIM 4
DIM 1	1,000			
DIM 2	,921	1,000		
DIM 3	,207	,113	1,000	
DIM 4	,129	,084	,645	1,000

Matriz de similaridades

Los índices más próximos son la Dimensión 1 ("Democracia de los ciudadanos") y la Dimensión 2 ("Democracia de las instituciones"), con un coseno de ,92. Al igual que con el coeficiente de correlación, las dimensiones 2 ("Democracia de las instituciones") y la dimensión 4 ("Democracia económica") son las menos similares, con un coseno de ,084, en la medición del Desarrollo Democrático.

	Uruguay	Costa Rica	Chile	Argentina	Bolivia	Perú	Nicaragua	Ecuador
Uruguay	1							
Costa Rica	0,97	1						
Chile	0,95	0,98	1					
Argentina	0,87	0,95	0,97	1				
Bolivia	0,82	0,87	0,84	0,87	1			
Perú	0,95	0,97	0,99	0,96	0,83	1		
Nicaragua	0,83	0,80	0,76	0,73	0,93	0,77	1	
Ecuador	0,93	0,96	0,96	0,95	0,94	0,96	0,896	1

Matriz de similaridades

Coseno de vectores de valores

Como podemos observar, se aprecian cambios en los coeficientes de similaridad cuando comparamos el coeficiente de correlación y el coseno. Esto suce-

de debido a que *el coeficiente de correlación equivale al coseno del ángulo entre los vectores cuando las variables se encuentran centradas*. Por ello, las diferencias que observamos entre el coseno y el coeficiente de correlación proceden del hecho que el coseno emplea los valores de los datos originales, con desviación expresada respecto al origen, mientras que el coeficiente de correlación emplea las variables normalizadas y con las desviaciones expresadas respecto a la media. Este hecho nos da alguna orientación sobre cuándo es preferible uno u otro coeficiente. *Cuando los datos tienen un origen con un significado claro (un cero como ausencia absoluta de alguna característica, por ejemplo), de forma que los valores tienen sentido, el coseno es la mejor opción para determinar la proximidad. Por el contrario, si el origen de los datos es arbitrario, y no tiene un significado concreto, expresarlos respecto a la media puede ser lo más conveniente, y por ello es recomendable emplear el coeficiente de correlación.*

Otras medidas empleadas son la covarianza y en general aquellas que expresan asociación.

2.1.2. *Datos expresando nivel de medición ordinal y rangos*

Al igual que el coeficiente de correlación es utilizado como indicador de similitud, pueden emplearse las versiones de correlación desarrolladas para variables ordinales y de rango, como son Spearman o Gamma, por ejemplo. Generalmente, las medidas de similitud basadas en la correlación como son Pearson, Mu2, Spearman, Gamma o Tau no se ven afectadas por las diferencias en las escalas de medición que se empleen en las variables.

Gamma

Se aplica cuando las variables son de tipo ordinal o rangos. Se calcula restándole a 1 el coeficiente de correlación gamma de Goodman-Kruskal. Su lógica es semejante a la del coeficiente de correlación. Como podremos observar, partiendo de la idea de emplear asociación como distancia o similitud, todos los coeficientes son susceptibles de ser empleados para tal fin. Es el caso de Spearman, o tau-b y tau-c, si bien estas últimas tienen dificultades para alcanzar los límites -1 y +1 cuando no se trata de tablas cuadradas.

2.1.3. *Datos expresando nivel de medición recuentos*

Cuando las variables expresan frecuencias también es factible determinar una medida de distancia entre ellas, tomando todas las categorías en conjunto. Las dos medidas de distancia más empleadas son el Chi-cuadrado y Phi-cuadrado.

a) Distancia

Medida de Chi-cuadrado

Un procedimiento frecuente para medir la disimilaridad cuando se emplean frecuencias se basa en Chi-cuadrado. El test de Chi-cuadrado determina si dos variables son o no independientes estadísticamente. Es decir, que no existe relación entre ellas. En su empleo como medida de distancia o disimilaridad, simplemente se calcula el coeficiente chi-cuadrado de la tabla de contingencia y se extrae su raíz cuadrada. En definitiva, esta medida de distancia consiste en la raíz cuadrada de chi-cuadrado. Cuando consideramos muchas variables, para cada par podemos estimar su distancia según chi-cuadrado, construyendo una matriz de disimilaridad entre las variables. Es importante recordar que las tablas son del tipo 2 x n ó n x 2, es decir que la comparación se continúa haciendo por pares, sean definidos por las filas (2) o por las columnas.

Dado que el valor que adquiera chi-cuadrado depende del tamaño muestral, la magnitud que alcance este coeficiente de distancia dependerá del número de casos considerados. Para intentar normalizar los coeficientes de distancia en variables que adoptan valores de frecuencia, se utiliza como alternativa el coeficiente phi-cuadrado.

Medida de phi-cuadrado

Esta medida intenta corregir el efecto que tiene la muestra en el cálculo del Chi-cuadrado. Para ello, divide la medición anterior, es decir, la medición de disimilaridad basada en chi-cuadrado, por la raíz cuadrada de las frecuencias totales (el total de casos u observaciones contenidas en la tabla de contingencia). Con ello, el valor del índice no se ve influido por las diferencias de frecuencias de las variables que se comparan.

Otros índices son la V de Cramer, el coeficiente de contingencia, Lambda y varios más.

2.1.4. *Datos expresando nivel de medición binario*

En el caso de datos binarios existen numerosos coeficientes de similaridad. Se consideran datos binarios los que se codifican con solamente dos valores. Ejemplos de esto pueden ser poseer una casa o no, estar sano o enfermo, hombre o mujer, etc. Si las observaciones son países, por ejemplo, las variables binarias pueden considerar poseer o no un sistema de salud pública, tener o no tener libertad de prensa, realizarse o no elecciones libres, etc. Habi-

tualmente la presencia de la característica o atributo se codifica con 1, y con un 0 su ausencia. En principio las diferentes mediciones de disimilaridad con datos binarios dependen de la importancia que le conceden a las diferentes celdas en la tabla de dos por dos que definen dos variables binarias. Consideremos como ejemplo dos casos y sus valores binarios en 5 variables. El código 1 indica "Sí" y el código 0 expresa "No". La estructura de la matriz de datos es rectangular, con los casos en filas y las variables en columnas. Este tipo de matriz es característica de las encuestas de opinión, así como de datos secundarios.

Tabla. Ejemplo A distancia entre casos: datos binarios

	Variable 1 *(tiene móvil)*	Variable 2 *(tiene TV)*	Variable 3 *(tiene radio)*	Variable 4 *(lee prensa)*	Variable 5 *(debate con amigos)*
Caso 1	1	1	0	0	1
Caso 2	0	1	0	1	0

Si consideramos las características que poseen en común y las que no, obtenemos una tabla de contingencia de 2 x 2. El caso 1 y el caso 2 coinciden que "sí" en la variable 2 (tener TV). Es decir, 1 coincidencia en que "sí-sí". Los dos casos coinciden que "no" en la variable 3 (tener radio). Es decir, 1 coincidencia en que "no-no". El caso 1 dice "sí" en dos ocasiones que el caso 2 dice "no" (variables 1 y 5). Es decir, 2 veces. Por último, el caso 1 dice "no" cuando el caso 2 dice "sí", en 1 sola ocasión, (variable 4).

Tabla. Ejemplo A distancia entre casos: cuadro resumen

		Caso 1	
		Sí	No
Caso 2	Sí	1	1
	No	2	1

En este ejemplo se compara las respuestas dadas por dos casos a las cinco variables. Partiendo de esta tabla cruzada, es posible estimar varios índices de similitud y disimilitud.

Al igual que en la ocasión anterior, puede efectuarse la misma operación para comparar dos variables (considerando los valores 0 y 1 presentes en los diferentes casos). Con ello construiremos una tabla comparando dos variables

dicotómicas. Por ejemplo, consideremos si las ciudades consideradas tienen una emisora propia de radio o televisión. Para determinar la proximidad o distancia en el equipamiento de los dos medios, construiríamos una tabla de siguiendo el mismo procedimiento anterior.

Tabla X. Ejemplo B distancia entre variables: datos binarios

	TV	Radio
Ciudad A	1	1
Ciudad B	0	0
Ciudad C	0	1
Ciudad D	1	0
Ciudad F	1	1
Ciudad G	1	1
Ciudad H	0	1
Ciudad I	1	0

Resumida según sus coincidencias y no coincidencias obtendríamos la tabla siguiente.

Tabla X. Ejemplo B distancia entre variables: cuadro resumen

		Radio		
		si	no	Total
TV	si	3	2	5
	no	2	1	3
Total		5	3	8

En cada una de las celdillas encontramos el número de ciudades que poseen o no las características consideradas. El total de ciudades, se expresa en la esquina inferior derecha.

La importancia que se le conceda a las diferentes casillas en la tabla depende de la naturaleza de las variables consideradas. Por ejemplo, consideremos cuando dos casos responden "no" a la pregunta ¿ha ganado alguna vez la lotería? Ese "no" aporta poca información con respecto a la similaridad entre los individuos. Lo más habitual es no ser agraciado con un premio, por lo que

la similitud con otro caso en la opción “no” aporta poca información. Por el contrario una respuesta positiva a esa pregunta puede indicar un parecido importante entre los dos casos. Los dos han sido premiados en un sorteo y esa coincidencia es algo que puede considerarse destacable. Sin embargo consideremos la característica “posee una emisora de televisión” y “posee una emisora de radio”. Para un país desarrollado, las coincidencias o parecidos en la respuesta negativa (no tener televisión o una emisora de radio) puede ser mucho más significativa que las coincidencias en la positiva. Es evidente que en cada situación se desea dar una importancia diferente a las coincidencias negativas o a las coincidencias positivas. En la primera queremos dar una mayor relevancia a las coincidencias de tipo "sí" (a los dos les ha tocado lotería), dada la rareza de la coincidencia. Es la misma situación que cuando los dos casos (ciudades) coinciden en la respuesta "no" (no poseen televisión o emisora de radio). Esta coincidencia en no tener “canal de televisión” y “emisora de radio” puede ser significativa (en algunos países) respecto a la similaridad entre los dos casos, y posiblemente más interesante que la coincidencia en "sí" tenerla.

Como ya hemos dicho, las medidas de similaridad en el caso de variables binarias se diferencian en el tratamiento que le dan a cada una de las casillas en la tabla que se forma. Algunas peticiones simplemente excluyen las casillas que expresan la ausencia de valor, es decir negativas, "no". En otras mediciones tendrán más peso las coincidencias que las diferencias, mientras que en otras se focalizan más en las diferencias que en las coincidencias. Evidentemente la selección de *la medida apropiada debe depender de la naturaleza de las variables y de la información que facilitan al investigador*. Es el investigador el que decide que características son más substantivas para los objetivos de su investigación.

La construcción de una tabla de doble entrada con dos variables binarias, define otra tabla de 2x2, cuyas celdillas notaremos con letras, según la combinación de presencia o ausencia de la característica. Estas letras van a ser usadas en la explicación de las medidas de disimilitud con datos binarios.

		Variable 1		
		1 (Sí)	0 (No)	Totales
Variable 2	1(Sí)	a	b	a + b
	0 (No)	c	d	c + d
	Totales	a + b	b + d	m= a+b+c+d

En la anterior tabla se tiene:

1. Donde *a* representa el número de individuos que toman el valor 1 en las dos variables de forma simultánea.
2. Donde *b* indica el número de individuos de la muestra que toman el valor 1 en la variable 2 y 0 en la variable 1.
3. Donde *c* es el número de individuos de la muestra que toman el valor 0 en la variable 2, y 1 en la variable 1.
4. Donde *d* representa el número de individuos que toman el valor 0 en las dos variables, al mismo tiempo.
5. Donde $a + c$ muestra el número de veces que la variable 1 toma el valor 1, independientemente del valor tomado por la variable 2.
6. Donde $b + d$ es el número de veces que la variable 1 toma el valor 0, independientemente del valor tomado por la variable 2.
7. Donde $a + b$ es el número de veces que la variable 2 toma el valor 1, independientemente del valor tomado por la variable 1.
8. Donde $c + d$ es el número de veces que la variable 2 toma el valor 0, independientemente del valor tomado por la variable 1.

Tomando como referencia la notación anterior, procedamos a la estimación de distancias y similitudes.

a) disimilitud

Distancia euclídea

Distancia euclidiana binaria. Tiene un *valor mínimo de cero y sin límite superior*.

$$\text{Distancia (x,y)} = \sqrt{(b+c)}$$

Se calcula a partir de una tabla 2x2 como la raíz cuadrada de $(b+c)$, donde b y c representan las casillas diagonales correspondientes a los casos presentes en un elemento pero ausentes en el otro.

Para el ejemplo A: $\sqrt{2-1+2} = 1{,}73$

Para el ejemplo B: $\sqrt{2-2+2} = 2$

Distancia euclídea al cuadrado

Tiene un valor mínimo de *cero y sin límite superior*.

$$\text{Distancia (x,y)} = b+c$$

Nuevamente, se calcula a partir de una tabla 2x2 como la suma de (b+c), donde b y c representan las casillas diagonales correspondientes a los casos presentes en un elemento pero ausentes en el otro.

Para el ejemplo A: 1 + 2 = 3

Para el ejemplo B: 2 + 2 = 4

Diferencia de tamaño

Se trata de un índice de asimetría. Tiene un *valor mínimo 0 y límite superior de 1*. Se calcula mediante $(b-c)^2 / n^2$. Siendo n el número total de casos.

$$\text{Distancia (x,y)} = (b - c)^2 / (a+b+c+d)^2$$

Para el ejemplo A: 1 / 25 = 0,04

Para el ejemplo B es cero dado que tanto b como c tienen el mismo valor.

Diferencia de configuración

Nuevamente b y c representan las casillas diagonales correspondientes a los casos presentes en un elemento pero ausentes en el otro, y a+b+c+d es el número total de observaciones al cuadrado. Su valor oscila de forma normalizada en un rango de cero a uno.

$$\text{Distancia (x,y)} = bc / (a+b+c+d)^2$$

Para el ejemplo A: 2/25 = 0,08

Para el ejemplo B: 4/64 = 0,063

Varianza

Se calcula a partir de una tabla 2x2 como (b+c)/4n, donde b y c representan las casillas diagonales correspondientes a los casos presentes en un elemento pero ausentes en el otro, siendo n el número total de observaciones. Oscila entre 0 y sin límite superior.

$$\text{Distancia (x,y)} = b+c / 4(a+b+c+d)$$

Para el ejemplo A: 3/20 = 0,15

Para el ejemplo B: 4/32 = 0,125

Forma

Esta medida de disimilitud, no tiene límite superior o inferior y penaliza la asimetría de las discordancias.

$$\text{Distancia (x,y)} = (a+b+c+d)(b+c) - (b-c)^2 / (a+b+c+d)^2$$

Para el ejemplo A: 14/25 = 0,56

Para el ejemplo B: 32/64 = 0,5

Lance y Williams

Se calcula donde *a* representa la casilla correspondiente a los casos presentes en ambos elementos y donde *b* y *c* representan las casillas diagonales correspondientes a los casos presentes en un elemento pero ausentes en el otro. Esta medida oscila entre 0 y 1. También se conoce como coeficiente no métrico de BrayCurtis.

$$\text{Distancia (x,y)} = b+c / 2a+b+c$$

Para el ejemplo A: 3/5 = 0,6

Para el ejemplo B: 4/10 = 0,4

b) Similitud[5] usando datos binarios

Como es habitual en las mediciones de asociación, una forma de medir la similaridad en variables dicotómicas es contar el número de veces que ambas variables toman el mismo valor de forma simultánea. La idea de referencia es que dos variables serán más parecidas cuantas mayores coincidencias se produzcan entre los valores de sus casos. Algo semejante a cuando empleamos el coeficiente de correlación anteriormente. Esto no obvia la necesidad de tomar varias decisiones importantes. Por ejemplo, qué hacer con las coincidencias 0-0, dado que si la dicotomía expresa la presencia o ausencia de una característica, la casilla *d* no tiene ningún significado real y cabria plantearse excluirla de la medida de similitud. La otra cuestión de interés es como ponderar las diagonales (las coincidencias y las no coincidencias) de la tabla de 2x2. Los índices de similitud que vamos a considerar son diferentes decisiones con respecto a las dos cuestiones anteriores.

Russell y Rao

La medición de similaridad de Russell y Rao se calcula dividiendo el número de coincidencias en la celdilla positivo-positivo, por el total de valores. Este coeficiente mide la probabilidad de que un individuo elegido al azar

5. Existen una multitud de índices de similitud para datos binarios como son: Rogers y Tanimoto, Sokal y Sneath 1, Sokal y Sneath 2, Sokal y Sneath 3, Kulczynski 1, Kulczynski 2, Sokal y Sneath 4, Hamann, Lambda, D de Anderberg, Y de Yule, Q de Yule, Ochiai, Sokal y Sneath 5, correlación Phi de 4 puntos, dispersión, etc.

tenga el valor 1 en ambas variables. Notemos que este coeficiente excluye la pareja 0-0, al contar el número de coincidencias pero no lo hace así al contar el número de posibles parejas. Asimismo, esta medida proporciona igual peso a las coincidencias y a las no coincidencias

$$\text{Distancia (x,y)} = a \,/\, a+b+c+d$$

Para el ejemplo A: 1/5 = 0,2

Para el ejemplo B: 3/8 = 0,37

Concordancia simple

La medición de concordancia simple, se define como el número de coincidencias divididas por el número total de casos (o de variables). Este coeficiente mide la probabilidad de que un individuo elegido al azar presente una coincidencia de cualquier tipo, pesando de igual forma las coincidencias y las no coincidencias.

$$\text{Distancia (x,y)} = a+d \,/\, a+b+c+d$$

Para el ejemplo A: 2/5 = 0,4

Para el ejemplo B: 4/8 = 0,5

En este ejemplo B, tenemos cuatro coincidencias entre los dos casos considerando siete variables por lo que el coeficiente de coincidencias sería cuatro dividido entre siete, o 0.5.

Jaccard

La medición Jaccard, excluye la celda negativa-negativa tanto del numerador como del denominador. Esta medida mide la probabilidad condicionada de que un individuo elegido al azar presente un 1 en ambas variables. Las coincidencias de tipo negativo-negativo (*d*) se excluyen al considerarse no significativas en este índice.

$$\text{Distancia (x,y)} = a \,/\, a+b+c$$

Para el ejemplo B el valor sería de .429.

Dice

La medición Dice excluye la valores coincidentes 0-0 tanto del numerador como del denominador y le asigna un peso doble al valor de las coincidencias del tipo 1-1. Se puede ver este coeficiente como una extensión de la medida de Jaccard, aunque su sentido probabilístico se pierde.

$$\text{Distancia (x,y)} = 2a \,/\, 2a+b+c$$

En el ejemplo B tendría un valor de .600

Rogers-Tanimoto

Este coeficiente puede interpretarse como una extensión de la medida de concordancias simples, pesando con el doble valor las no coincidencias $(b+c)$.

$$\text{Distancia (x,y)} = a+d \,/\, (a+d+2(b+c)$$

En el ejemplo B tendría un valor de .333

Medida de Kulcynski

Esta medida es el cociente entre coincidencias y no coincidencias, excluyendo los pares negativo-negativo.

$$\text{Distancia (x,y)} = a \,/\, b+c$$

Medida Φ Phi

Al igual que se utiliza el coeficiente de correlación de Pearson como medida de proximidad, en el caso de tablas de 2x2 es posible emplear el coeficiente de Correlación Phi de 4 puntos. Este índice su equivalente en binario, con un rango de variación entre -1 y +1.

En el ejemplo B tendría un valor de -.067

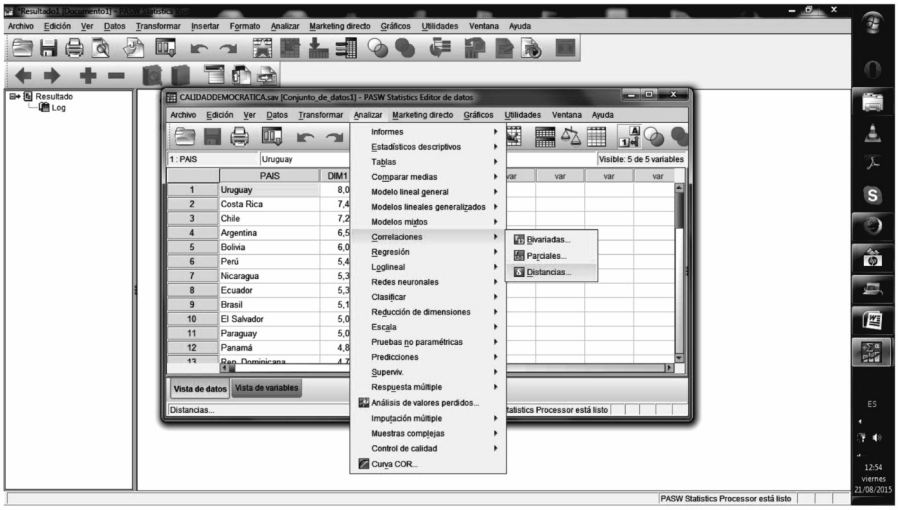
Vamos a continuación a considerar algunas de las aplicaciones inmediatas de las matrices de distancias o proximidades. Una de ellas es, evidentemente, el análisis de conglomerados. Siendo el concepto de proximidad y distancia una noción fundamental, en general, es una cuestión de interpretación su empleo con un sentido teórico u otro. Así, la asociación (como es el caso de la correlación) puede interpretarse como proximidad o distancia. Como similitud o disimilitud. En un sentido metafórico, los análisis de asociación, en especial los basados en modelos lineales, simplemente expresan la idea de que todas las variables son, hasta cierto punto y tras las modificaciones que producen las transformaciones de combinación lineal, un sistema de copias que reflejan con mayor o menor precisión las imágenes de unas en las otras.

La investigación social experimentó una revolución tras la implementación de paquetes informáticos que realizan tareas de análisis estadístico. Y cabe enfatizar, tareas de tratamiento de datos. Los programas no efectúan análisis en el sentido de interpretación. Es el investigador quien investiga y analiza. Es el investigador quien busca sentido en los datos que representan

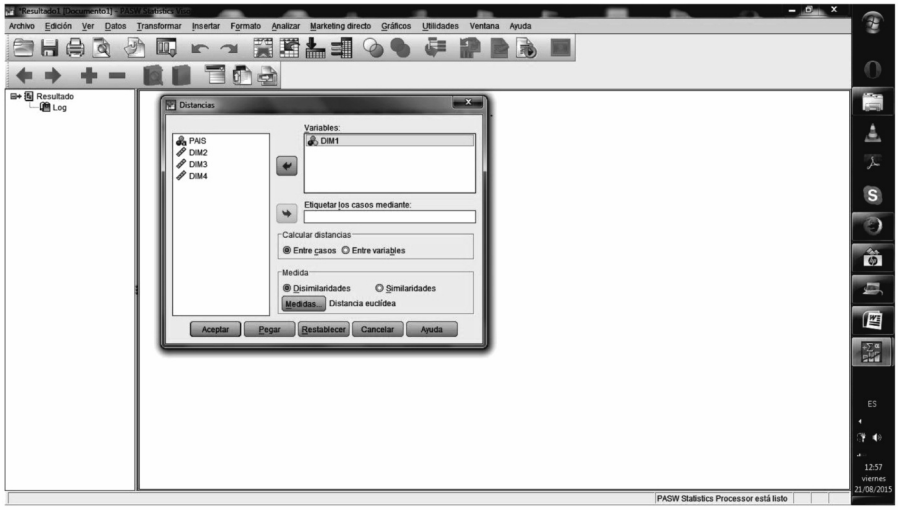
la realidad, con todas las limitaciones por todos conocidas. Los programas son una herramienta. Una ayuda valiosa que permite afrontar tareas que serían casi imposibles por su coste en tiempo para cualquier investigador. La oferta de programas comerciales es muy elevada y cada vez son más. Una opción interesante es utilizar programas liberados como son el programa R. Sin embargo, su curva de aprendizaje es lenta y exige una documentación extensa de procedimientos para aquellos que no están familiarizados con él. La paradoja está servida. Aquellos que conocen el programa, dada su especialización, posiblemente ya saben dónde encontrar estos análisis (incluso buscando en CRAN). Para los no competentes con R, obliga a escribir un manual formativo que excede este texto. Hoy por hoy, los programas comerciales son más intuitivos y fáciles de usar. Basta unas orientaciones básicas y la curva de aprendizaje es rápida. Obviamente, al ser un producto comercial que compite en un mercado en expansión, buscan la fórmula de hacerlos más acogedores. En esta ocasión los ejemplos se expondrán en dos programas bastante extendidos: SPSS y SYSTAT.

2.2. Ejemplos en SPSS y SYSTAT

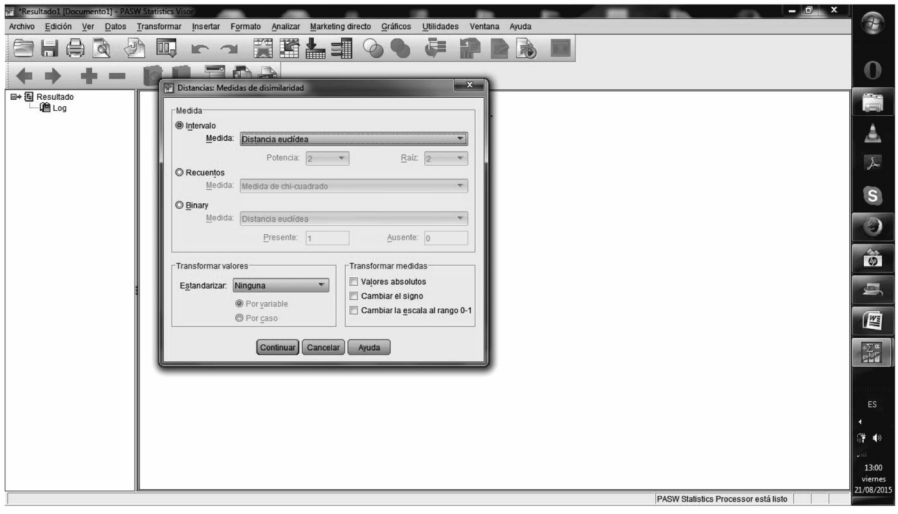
En el programa SPSS las distancias y las similitudes se obtienen desde el menú "Analizar", opción "Correlaciones" y dentro de ella "Distancias".



Una vez elegida la opción "Distancia", aparecen las dos ventanas, a la izquierda el listado de variables existente en la base de datos y a la derecha las variables que se eligen para calcular las distancias o similitudes.

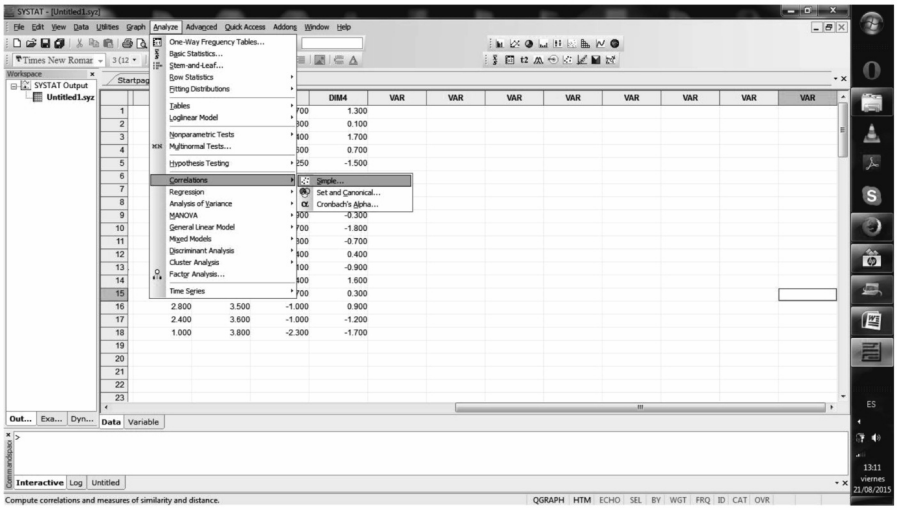


Es posible elegir las distancias ente casos o variables, así como que medida de distancia o similitud se desea calcular. Las medidas disponibles en SPSS se muestran desplegando la opción "medidas". Están organizadas según métrica de las variables en Intervales, Recuentos y Binarias. Al elegir el tipo de medida, se activa el desplegable de la derecha dando a elegir qué índice se desea emplear. Los índices serán del tipo elegido en la ventana anterior: de disimilaridad o de similaridad.



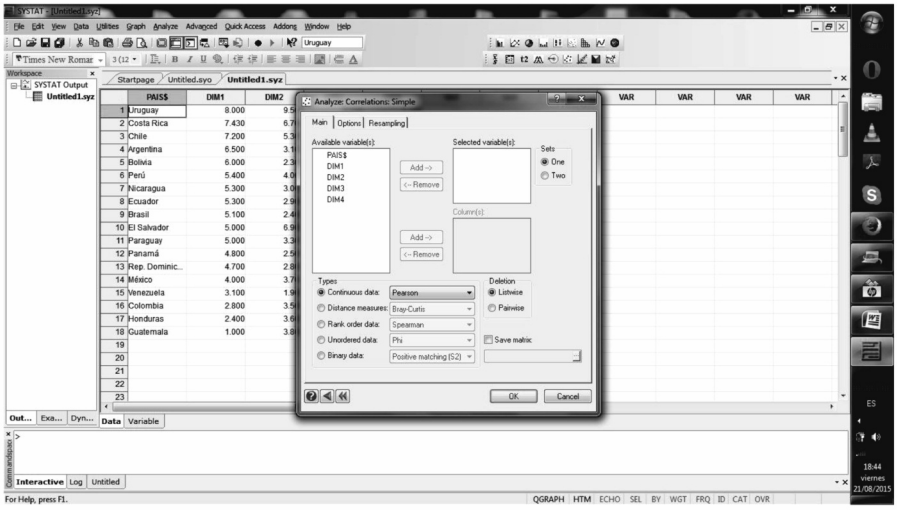
En la zona inferior izquierda de la ventana se muestran las opciones de transformar los valores, ya sea en fila (casos) o columnas (variables). A la derecha, la opción de normalizar los índices de similitud o disimilitud. Las transformaciones de casos o variables se utilizan para el cálculo de los índices de similitud o disimilitud.

En el programa SYSTAD los menús son bastante semejantes. El cálculo de las matrices de disimilitud y similitud (proximidad y distancia) se encuentran en la opción "Analizar", y dentro del desplegable la opción "Correlación" y nuevamente "Simple". El procedimiento para estimar las distancias se encuentra incorporado dentro del sistema de opciones de "Correlación".

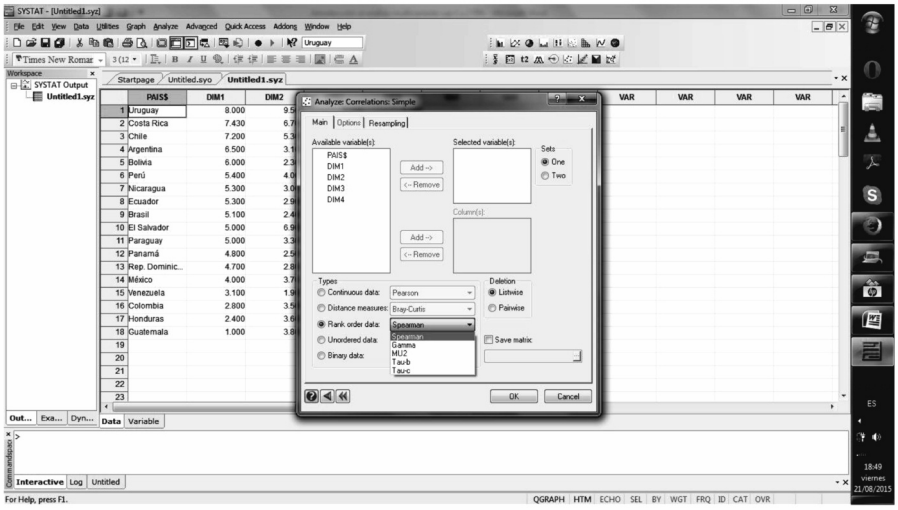


En el SYSTAD cabe la posibilidad de producir matrices de distancias cuadradas o simétricas. En el caso de elegir un solo grupo, se produce una matriz simétrica. Al elegir un grupo se desactiva la segunda ventana de selección de variables.

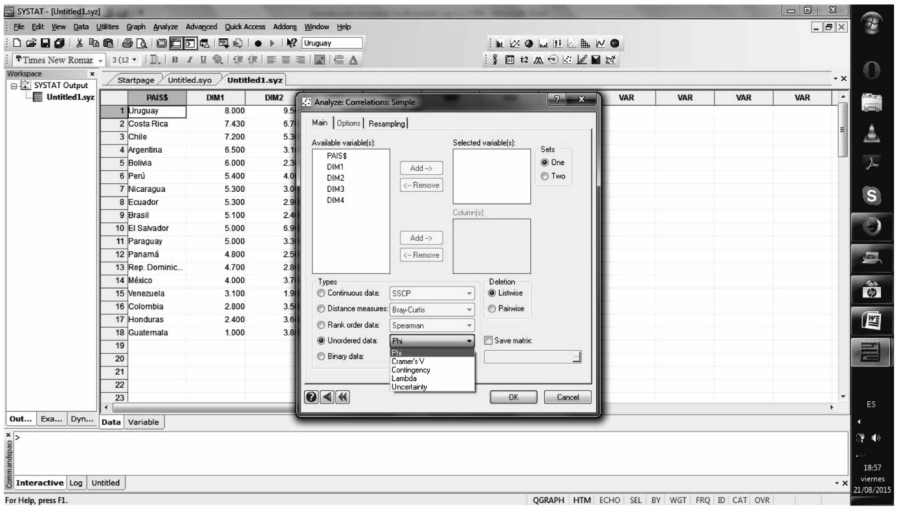
Las diferentes distancias están agrupadas según una lógica más detallada. En el área inferior izquierda se puede elegir entre Intervalo, datos expresando Distancias, medición ordinal y de rangos, datos categoriales o recuentos y binarios. Dentro de cada opción se encuentran una importante diversidad de indicadores de distancia.



Algunos ejemplos, las distancias para datos ordenados y las distancias para datos no ordenados. Para datos ordenados, como son las variables ordinales o los rankings.



Y para datos de tipo recuento, es posible emplear en SYSTAD los siguientes índices.



Las otras dos pestañas, Opciones y Remuestreo, se refieren a estrategias para estimar la significación. En el caso de las "opciones" solamente son utilizables con variables de intervalo.

3. EL ANÁLISIS DE CONGLOMERADOS

"Ódiame por piedad, yo te lo pido…
¡Ódiame sin medida ni clemencia!
Más vale el odio que la indiferencia.
El rencor hiere menos que el olvido.

Yo quedaré, si me odias, convencido,
de que otra vez fue mía tu existencia.
Más vale el odio a la indiferencia.
¡Nadie aborrece sin haber querido!"

El último ruego[1] (1903)
FEDERICO BARRETO (1862-1929)

Como hemos comentado anteriormente, el cálculo de una medida que estime la proximidad o distancia entre casos o variables, nos permite establecer una serie de procedimientos de análisis. En términos intuitivos, posiblemente el más inmediato se refiere a la posibilidad de formar grupos de casos a partir de la proximidad que se haya establecido entre ellos. El interés es evidente, en la medida que agrupar los casos, por ejemplo países, instituciones, individuos, grupos, asociaciones, nos permite establecer posibles tipos diferenciados, en función a las características que les hacen estar próximos. El procedimiento estadístico más generalizado que considera la agrupación de objetos o casos en función a su similitud o disimilitud es el denominado análisis de conglomerados.

De forma simple, el análisis de clúster consiste en identificar la existencia de grupos en los datos u observaciones. Así, para Kaufman y Rousseeuw

6. Barreto F. El último ruego. No. 35 de la revista *Actualidades*, nº. 35, 21 de septiembre de 1903, página 576

(1990) el análisis de clústeres es el arte de encontrar grupos en los datos. No obstante, definir qué es el análisis de clúster es una tarea bastante complicada, tanto por la gran variedad de métodos utilizados como por la presencia de una importante diversidad conceptual. Everitt et al. (2011, 7) revisando los conceptos de "clúster", "grupo" o "clase" proponen que intentar definir de forma única cada uno de esos conceptos, en el marco del análisis de clústeres, puede ser algo que genera más confusión que ayuda.

Uno de los rasgos de este tipo de análisis es su carácter exploratorio, en la medida que no es preciso conocer previamente ningún tipo de pertenencia o tipología para investigar las agrupaciones que forman los casos. Esta posibilidad de decidir cuántos son los grupos relevantes, es propia tanto de los modelos jerárquicos como de los no jerárquicos. Tomados un conjunto de casos, y a partir de ellos las matrices de distancias, es posible determinar cuáles son más similares entre sí y cuales más diferentes. Como ya sabemos, se pueden medir las matrices de distancias entre casos o también entre variables. Como era previsible, los métodos de conglomerados (clúster) permiten agrupar tanto casos (por ejemplo países) como variables (por ejemplo, indicadores de desarrollo). La finalidad es la misma: agrupar lo semejante y diferenciarse de lo diferente. Ese carácter exploratorio se evidencia en que muchos de los procedimientos de "minería de datos" se basan en el análisis de clústeres. Son muchos los autores que enfatizan dicha dimensión exploratoria. Everitt afirma como "Muchas de las técnicas de análisis de clústeres han ocupado un lugar junto a otras técnicas de análisis exploratorio de datos, entre las herramientas que emplean los estadísticos. El término exploratorio es importante aquí porque explica la ausencia del coeficiente "p-value", presente en muchas otras áreas de la estadística. (…/…) los métodos de clúster están pensados más para generar hipótesis que para testarlas". (1993, 10).

Existen otros métodos que proponen determinar de forma automática el número de clústeres existentes en la matriz de datos. Sin embargo, cuando se trata de investigación, todos los procesos que actúan alejando al investigador del contacto con los datos son claramente perjudiciales. Todo análisis consiste en una sucesión continuada de toma de decisiones sobre la pertinencia y significado teórico de lo que se descubre en los datos. Operar sobre los datos, evaluando y diagnosticando el significado de las diferentes soluciones de número de clústeres permite un mejor conocimiento de sus significados y condiciones en que se estudia su estructura. En ese sentido, las facilidades de análisis, en el sentido de ofrecer un número concreto de clústeres, esconden y lo que es aún peor, permiten, ofrecer resultados sin conocer las limitaciones.

Los procedimientos para definir un clúster pueden permitir que estos se solapen, por lo que un caso puede pertenecer a dos o más clústeres simultáneamente o permitir solamente la pertenencia a un clúster. Este último es el procedimiento más generalizado. Como es habitual, los resultados que obtengamos dependerán directamente de la información que hayamos utilizado para efectuar el análisis. Esto significa que tanto el número de conglomerados como las diferencias que se establecen entre estos grupos, significarán una cosa u otra en función a las variables que se hayan utilizado para determinar las diferencias o similitudes. La consecuencia es evidente. Un grupo de objetos o casos, por ejemplo países, formarán grupos o tipos diferentes en función a las variables que se tomen en consideración para definir la matriz de distancias o proximidad. Evaluar la pertinencia es fundamental para conocer el significado real de los grupos o tipos que se extraen. En todo caso, es importante insistir en que la selección de variables es esencial para los resultados que se obtengan, especialmente el significado que se les puede atribuir.

Asimismo, además del significado sustantivo en términos teóricos que pueda tener el utilizar unas u otras variables de referencia, existen varias decisiones de carácter técnico que condicionarán los resultados que obtengamos. Por ejemplo, qué medida de distancia o proximidad se elija, o los criterios que decidirán la adscripción de un caso a un conglomerado u otro (por ejemplo, qué decidir cuando un caso se encuentre a una distancia similar de dos grupos). Como ya sabemos, las dos medidas (de distancia o de proximidad) son realmente dos formas de mirar un mismo fenómeno. Por ello, los métodos siguientes de agrupación son aplicables en los dos tipos de medición, simplemente aplicando la misma lógica de forma inversa. Es decir, tomar disimilitud por similitud.

Respecto a la cuestión sobre qué casos son más similares o más diferentes, la respuesta la encontramos recurriendo a las medidas de similitud o disimilitud que hemos presentado en el capítulo anterior. Todo aquello que se comentó respecto a las limitaciones en el uso de unas medidas u otras, se aplica de forma directa a los procedimientos estadísticos que se apoyan sobre ellas para poder operar. Los procedimientos de detección de clústeres más empleados son las técnicas jerárquicas y las que operan sobre K-medias. Estos parten de la ventaja de haber sido muy testados y sus limitaciones bien conocidas (Bacher 2000: 223; Everitt et al. 2001: 94-96; Huang 1998: 288).

3.1. Métodos de vinculación, distribucionales y de densidad

De igual modo que existen muchas formas diferentes para calcular la distancia entre casos u objetos, también existen varias alternativas para poder combi-

narlos en diferentes grupos. Éstas alternativas siguen diferentes lógicas, especialmente en las situaciones donde el número de casos que deben ser agrupados es muy elevado. Evidentemente, los procedimientos para agrupar un número limitado de casos permiten unas herramientas analíticas diferentes a cuando son muchos casos. Vamos a considerar dos grandes procedimientos: los métodos de aglomeración jerárquicos y los no jerárquicos.

En los métodos jerárquicos, un clúster contiene otros clústeres, quienes a su vez contienen otros clústeres y así sucesivamente hasta finalizar en un solo grupo agrupando todos los clústeres. Es un procedimiento que opera tanto de forma inclusiva agregando clústeres como divisiva, separándolos progresivamente. Por lo general, los métodos jerárquicos son plenamente exploratorios, de forma que decidir cuántos grupos existen es el resultado del análisis. Los métodos jerárquicos utilizan sobre todo procedimientos de vinculación (linkage) entre casos o variables. En ese sentido, los métodos jerárquicos permiten establecer clústeres de variables, de casos o de ambos simultáneamente.

En los métodos no jerárquicos, lo más frecuente es ir decidiendo cuántos grupos al iniciar el análisis (tal y como sucede con el análisis discriminante). Con este enfoque, solamente se establecen grupos de casos, no de variables. Los clústeres son el resultado de la partición en grupos de los casos en estudio. En definitiva, en el primer caso los grupos se van agregando incrementando la heterogeneidad del grupo resultante, mientras que en los métodos de partición, como son k-medias o k-medianas, se separan los casos intentando optimizar las diferencias entre los grupos solicitados y buscando la mayor homogeneidad interna.

Considerando estos dos tipos de procedimientos para buscar clústeres (jerárquicos y k-clústeres), ya existe una gran diversidad de métodos diferentes. La mayoría de los métodos permiten elegir qué medida de similitud o disimilitud emplear para formar los grupos. De hecho, existe una inmensa lista de opciones de medidas de similitud y disimilitud. Por ejemplo, si consideramos Minkowski, es muy elevado el número de medidas de distancia que podemos definir. Para incrementar aún más la diversidad, aparece la opción, que ya hemos considerado, de transformar las variables (mediante normalización o estandarización). Otra cuestión relevante son las reglas que podemos establecer para decidir cuántos son los grupos existentes en los datos. Estas reglas o criterios para dar por finalizada la exploración son más abundantes de lo que pudiese parecer, llegando Milligan y Cooper (1985) a identificar y analizar hasta 30 reglas o criterios diferentes. Si combinamos todas las

opciones anteriores (tipos de análisis, métodos de análisis, medida de similaridad o disimilaridad elegida, trasformaciones y reglas de finalización), podemos apreciar la gran cantidad de tipos de análisis existentes.

La confusión se agrava por el hecho de que diferentes disciplinas han producido, para sus análisis métodos muy parecidos para el análisis de clústeres, si bien les han denominado de formas diferentes. En el caso de los métodos jerárquicos, "hierarchical clustering" (McQuitty, 1960; Johnson, 1967); "single linkage clustering" (Sokal and Sneath, 1963), o "joining" (Hartigan, 1975). En lo referido a los procedimientos para producir las agregaciones (o desagregaciones según se elija) de los grupos, Blashfield y Aldenderfer (1978) facilitan una relación de equivalencias entre los términos empleados, que Jain y Dubes (1988), y Day y Edelsbrunner (1984) completan incluyendo sinónimos y acrónimos. La relación siguiente muestra varias de esas equivalencias y sinónimos. Un investigador, según su disciplina, tenderá a emplear unos u otros términos. No obstante, se referirán en la práctica al mismo procedimiento. En este texto mantenemos los nombres en su versión inglesa, en un intento de reducir la confusión ya existente, y que se incrementa aún más cuando median traducciones.

Sinónimos y equivalencias

Single linkage: Nearest-neighbor method, Minimum method, Hierarchical analysis, Space-contracting method, Elementary linkage analysis, Connectedness method.

Weighted average linkage: Weighted pair-group method using arithmetic averages, WPGMA, Weighted group-average method.

Centroid linkage: Unweighted centroid method, Unweighted pair-group centroid method, UPGMC, Nearest-centroid sorting

Complete linkage: Furthest-neighbor method, Maximum method, Compact method, Space-distorting method, Space-dilating method, Rank-order typal analysis, Diameter analysis.

Median linkage: Gower's method, Weighted centroid method, Weighted pair-group centroid method, WPGMC, Weighted pair method, Weighted group method.

Average linkage: Arithmetic-average clustering, Unweighted pair-group method using arithmetic averages, UPGMA, Unweighted clustering, Group-average method, Unweighted group mean, Unweighted pair-group method.

Ward's method: Minimum-variance method, Error-sum-of-squares method, Hierarchical grouping to minimize tr(W), HGROUP.

Los nombres en negrita serán los utilizados en este libro cuando nos refiramos y describamos los métodos para establecer la vinculación entre clústeres.

Como hemos observado anteriormente, el análisis de clúster es una estrategia fundamental en las tareas de minería de datos y en lo que actualmente se denomina "Big data". Es decir, de la explotación exploratoria de grandes bases de datos que incorporan información de diverso tipo. Este hecho hace que partiendo de los métodos anteriores se hayan desarrollado otros procedimientos alternativos. Es el caso de la búsqueda de clústeres basándose en las *distribuciones multivariables* o los que se basan en el *estudio de la densidad*. Las estimaciones de densidad (Hartigan 1975; Wong and Lane, 1983) pueden ser aplicadas al análisis de clústeres, existiendo varios métodos para ello (Silverman, 1986).

El método para detectar clústeres más directamente conectados con la estadística es el basado en el análisis conjunto de distribuciones. Para ello se modifica levemente la definición de clúster de forma que incluya el concepto de distribución. Un clúster estará formado por aquellos casos que con mayor probabilidad pertenezcan a una misma distribución. Este tipo de modelado presenta el problema del sobreajuste, de forma que el investigador debe establecer una serie de criterios y condiciones que limiten las soluciones posibles de los modelos. Por definición, cuanto más complejo es un modelo mejor ajustará sobre la diversidad de los datos, sin embargo la noción misma de parsimonia ya sugiere que el modelo más complejo no es necesariamente el mejor, aun cuando sea el más explicativo. En esta lógica distribucional de la exploración de clústeres, los clústeres capturan y expresan la correlación e interdependencia entre los atributos empleados para agrupar los casos. Entre los métodos más empleados se encuentran los modelos gaussianos mixtos, dónde los datos son modelados mediante un número fijo de distribuciones gaussianas, inicializadas mediante valores aleatorios, y que mediante un procedimiento de ajuste iterativo busca optimizar su ajuste a los datos. Los casos se atribuyen a aquella distribución gaussiana a la que es más probable que pertenezcan. Como es habitual en este tipo de ajustes, el riesgo de un mínimo local (con lo que el ajuste no sería óptimo considerando toda la función) siempre está presente, por lo que se habitúa a efectuar varios intentos, en los que pueden encontrarse soluciones diferentes.

Los métodos que emplean el enfoque del análisis de la densidad, plantean que los clústeres vienen definidos por áreas donde los casos se concentran densamente. Estás áreas de concentración de casos estarían rodeadas de zonas de baja densidad, que delimitarían a los diferentes clústeres entre sí. Los casos presentes en esas zonas difusas son definidos como "ruidos" o casos

atípicos. Partiendo de esta idea, existen diferentes métodos para darle una forma operativa. Algunos de ellos aún no están incorporados en la mayoría de los programas comerciales más extendidos, al surgir asociados a la minería de datos y el análisis de Big data. El método DBSCAN propone un modelo de conglomerado basado en el alcance de densidad (density-reachability), y consiste en conectar aquellos casos que se encuentran espacialmente dentro de un intervalo. Para poder incluir los casos dentro del clúster estos deben cumplir unos criterios de densidad (como es un número mínimo de casos dentro de un determinado radio), por lo que el clúster consiste en todos los casos que están densamente conectados más todos los casos que se encuentran dentro de un radio de influencia de ese grupo. En ese sentido, los clústeres pueden adoptar formas muy irregulares. Este método emplea dos parametros: ε (el rango de influencia que "atrapa" a los casos dentro del clúster) y el número mínimo de puntos (minPts) a partir del cual podemos concluir que existe una región especialmente densa y definitoria de un clúster. Por lo general, el procedimiento se inicia con un caso cualquiera y se determina si el número de casos que captura el parámetro ε es suficiente para definir un clúster. En caso afirmativo se identifican como un clúster. Todos los casos que forman parte de ese clúster incorporarán, a su vez, al clúster aquellos otros casos que se encuentren dentro de su radio de influencia ε. El proceso continúa hasta "cerrar" el clúster dado que todos los demás casos estarían fuera del área de influencia. Entonces se reinicia nuevamente el proceso comenzando con otro caso fuera del clúster, en búsqueda de posibles nuevos grupos. Cuando el caso de inicio no define un clúster es etiquetado provisionalmente como "ruido", aunque más tarde pueda ser capturado dentro del radio de influencia de otro clúster y con ello ser incorporado a él. Otras variantes de este método son OPTICS (que elimina la necesidad del parámetro ε mediante la generación de clústeres jerárquicos) o DeLi-Clu (Density-Link-Clustering) que combina los métodos de linkage simple con OPTICS. Es evidente que estos métodos tienen limitaciones, muchas de ellas equivalentes a todos los que se basan en el concepto de distancia. La calidad del resultado depende, como ya sabemos, de la distancia elegida. La más habitual en DBSCAN es la distancia euclidea, que siendo una medida de distancia eficaz, presenta serios problemas cuando existe una elevada dimensionalidad en los datos. Esa hace muy dificultosa la tarea de decidir un valor apropiado para ε. Recordemos, asimismo, el efecto de las métricas que se empleen, y que también afectará al radio de influencia. Otra limitación es cuando los clústeres muestran grados diferentes de compactación. Si son muy desiguales

en su densidad, elegir un número mínimo necesario para definir un clúster y que sea válido para todos ellos, se complica seriamente. Otro elemento de dificultad es el tratamiento de las zonas "difusas" formadas por casos que no pertenecen a ningún clúster y que finalmente pueden incorporarse a unos u otros según el orden de ejecución. Así, un caso atípico puede incorporarse a un clúster, simplemente porque se definió primero, pudiendo sin embargo formar parte de cualquier otro clúster construido posteriormente. Dimensionalidad en los datos, métrica, heterogeneidad en la densidad de los clústeres o tratamiento de los casos pertenecientes a las áreas difusas en los bordes de los clústeres son algunos de los problemas a gestionar en estos métodos de análisis.

Otro de los métodos para agrupar los casos según la densidad emplea la estimación de densidad kernel. No nos extenderemos en detalle en el procedimiento por razones de espacio. La estimación de densidad kernel es un procedimiento no paramétrico para estimar la función de densidad de una variable aleatoria. En definitiva, una estimación de la probabilidad de que la variable aleatoria adopte un valor dado. Un ejemplo de esta estrategia de conglomeración es desplazar los casos hacia las áreas más densas, basándose en la estimación de densidad kernel. Los casos convergerían en un máximo local de densidad, y esos "atractores" de densidad actuarían como representación de los clústeres. Otros dos métodos de agrupación basados en la densidad son el Kernel uniforme y el k_{th} vecino más próximo (k_{th} nearest neighborhood). Los dos métodos calculan una estimación de valor para cada caso basada en la función de densidad. Partiendo de esa nueva estimación y de la matriz de disimilaridad original se construye una nueva matriz de disimilaridad. Finalmente, se aplica el método de linkage simple empleando la nueva matriz de disimilaridad. En el procedimiento de kernel uniforme, se facilita un valor para el radio r. La densidad de un caso x se calcula como la proporción de casos incluidos en la esfera de radio r y centrada en ese caso x. En el procedimiento k_{th} vecino más próximo, se facilita un valor para k, y a partir de él, se calcula la densidad del caso x como la proporción de casos incluidos en la esfera con centro en el caso x y con radio la distancia al caso vecino más próximo al valor k. En los dos métodos, la estimación de la nueva medida de disimilaridad para dos casos se calcula como la media de los valores de densidad de los dos casos, siempre y cuando los dos casos se encuentren dentro de la misma esfera de referencia. El desarrollo de estrategias de conglomeración está en clara expansión dada su importancia en la investigación de minería de datos y Big Data.

3.2. Métodos jerárquicos

"¿Que vales más que yo? ¡Tonta, orgullosa!
Vales lo que tu carne blanca y dura.
Los dos, al fin, entre una humilde fosa
vestiremos la misma vestidura.

Y cuando yo descanse en el osario,
fatigado del mundo y del perverso,
papeles revolviendo un anticuario,
quizá encuentre mi nombre al pie de un verso.

Mientras tú, que en la lid de la existencia
palma de vencedor has obtenido
después de un de que indica pertenencia
llevarás a lo sumo otro apellido".
A ***[7]

Guillermo Valencia Castillo (1873-1943)

Cuando el número de casos no es excesivo, una de las lógicas más frecuentes es establecer un procedimiento jerárquico. Los métodos jerárquicos son generalmente de dos tipos, según se parta de tantos grupos como casos, o considerando todos los casos como un clúster y posteriormente ir desagregando. El primer procedimiento se denomina por aglomeración y el segundo por división.

Los métodos por aglomeración comienzan considerando cada caso como un grupo separado. En definitiva, N grupos con un tamaño de 1. Los dos grupos (casos) más próximos se unen en un único clúster. En ese momento existirán N-1 grupos, con uno de ellos de tamaño 2 y el resto de tamaño 1. Este procedimiento continúa hasta que todos los casos pertenecen a un único grupo. Así, partiendo de los casos individuales, los va agrupando sucesivamente de forma que finalmente en un último paso definen un único grupo. En la ejecución de este proceso, los casos forman grupos que a su vez son agregados en otros grupos en un proceso de simplificación. Al inicio existen tantos grupos como individuos, en un segundo paso, dos casos forman un grupo. En un tercer momento, o bien un nuevo caso es agregado a este grupo o dos casos diferentes se unen formando un segundo grupo. Conforme el proceso

7. Dra. María Helena Barrera Agarwal, quien el domingo 3 de julio del presente año publicó el artículo "Los orígenes de Odiame" en la revista Artes del diario ecuatoriano La Hora. Dicho artículo se encuentra en la página 7 de la mencionada revista y puede ser leído en la siguiente dirección: http://issuu.com/la_hora/docs/artes030711

de agrupación va avanzando, a) nuevos casos se incorporan a grupos ya existentes, b) definen ellos mismos un nuevo grupo, o c) se unen en un solo grupo otros grupos ya preexistentes. Una característica de los métodos jerárquicos es que una vez asignado un caso a un grupo, ya no puede ser retirado de él, como tampoco pueden subdividirse grupos ya existentes. Como puede apreciarse, los clústeres se van agrupando de forma jerárquica, donde el superior engloba a otros más pequeños.

Los métodos de tipo divisivo comienzan con todos los casos formando un único grupo. Este grupo se divide según el criterio que se decida para crear dos grupos. Posteriormente uno de esos dos grupos se divide en otros dos, de forma que se generan tres grupos. Nuevamente, uno de los tres grupos se subdivide para formar otros dos, produciendo un total de cuatro grupos. Se continúa hasta que finalmente hay tantos grupos como casos. Si bien es un procedimiento alternativo a los procedimientos por conglomeración jerárquica, son bastante infrecuentes tanto en aplicaciones concretas como en opción de análisis en la mayoría de los programas de análisis. Las dos estrategias para generar los grupos son bastante exigentes desde el punto de vista estadístico al implicar múltiples comparaciones. Como observan Kaufman y Rousseeuw (1990), en el primer paso de cualquier procedimiento jerárquico aglomerativo se deben considerar N(N-1)/2 pares de observaciones o casos a efectos de determinar cuáles son los más similares. El número de pares crece exponencialmente conforme crece el valor de N (número de casos u observaciones). En los procedimientos de formación de clústeres mediante división, el primer paso es elegir los dos subgrupos (no vacios) que menos se parecen (más disimilares). Considerando todas las posibilidades, implica $2^{(N-1)} - 1$ comparaciones. Al igual que en el procedimiento anterior, el número de comparaciones crece de forma exponencial conforme crece N.

En todo caso, es el investigador quien decide qué criterio (similaridad o disimilaridad) se va a emplear para fusionar los casos en un clúster. Cuando en un grupo hay más de un caso, debe decidirse qué criterio se va a seguir para determinar si los grupos son más o menos próximos (similares). Estos procedimientos para comparar grupos se denominan métodos de vinculación (linkage methods). La definición del término "más próximo" es diferente para cada método de vinculación (linkage). Por ello, dependiendo del método empleado, la matriz de distancias (o disimilaridad) que se obtine después de cada fusión se calcula mediante fórmulas diferentes. Al comenzar el proceso, se emplea la matriz de distancias original, pero esta varía conforme se van produciendo las agregaciones de clústeres. Esa es la diferencia clave entre métodos: como se calcula la nueva matriz de distancias cada vez que se fusionan dos grupos.

Lance y Williams (1967) desarrollaron una fórmula que permite considerar, como casos especiales, la mayor parte de los métodos más conocidos de conglomeración jerárquica. Esta propuesta ha sido debatida por múltiples autores como Anderberg (1973); Jain y Dubes (1988); Kaufman y Rousseeuw (1990); Gordon (1999); Everitt et al. (2011); and Rencher and Christensen (2012), mostrando como los diferentes métodos de conglomeración pueden ser incluidos en ella. De acuerdo con la notación de Everitt et al. (2011, 78), la fórmula de Lance–Williams puede expresarse de la forma siguiente

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \ \left| d_{ki} - d_{kj} \right|$$

donde d_{ij} es la distancia o disimilaridad entre el clúster *i* y el clúster *j*; $d_{k(ij)}$ es la distancia entre el clúster k y el nuevo clúster formado al fusionar los clústeres i y j; α_i , $\alpha_j \beta$, γ son parámetros que varían según el método de vinculación (linkage) que se elija en cada caso. Esta fórmula recurrente permite que se pueda calcular la disimilaridad entre los nuevos grupos creados y los grupos existentes en cada fase de la agrupación jerárquica. Consideremos un ejemplo concreto para presentar los diferentes coeficientes α, β, γ. Pensemos que R, P y Q son clústeres existentes y que se va a generar un nuevo grupo resultado de la fusión entre P y Q. Este nuevo grupo vendrá notado por P+Q, y donde n_P nota el número de objetos (casos) existentes en el clúster P, n_R el número de casos en el clúster R y n_Q los casos en el clúster Q. La distancia entre los clústeres R y el nuevo formado por P+Q vendría planteada en la siguiente ecuación

$$d_{(R,P+Q)} = \alpha_1 d_{(R,P)} + \alpha_2 d_{(R,Q)} + \beta d_{(P,Q)} + \gamma \ \left| d_{(R,P)} - d_{(R,Q)} \right|$$

donde los parámetros varían según el método que se aplique, como puede apreciarse en la tabla siguiente. Las distancias son diferentes para cada método así,

Método de vinculación (linkage) $d_{(R,P+Q)} =$	α_1	α_2	β	γ
Single	1/2	1/2	0	–1/2
Complete	1/2	1/2	0	1/2
Average	$n_P / (n_P + n_Q)$	$n_Q / (n_P + n_Q)$	0	0
Weighted	1/2	1/2	0	0
Centroid	$n_P / (n_P + n_Q)$	$n_Q / (n_P + n_Q)$	$-(n_P n_Q / (n_P + n_Q)^2$	0
Median	1/2	1/2	–1/4	0
Ward	$(n_R + n_P) / (n_R + n_P + n_Q)$	$(n_R + n_Q) / (n_R + n_P + n_Q)$	$- n_R / (n_R + n_P + n_Q)$	0
Flexibeta	$(1 - \beta) / 2$	$(1 - \beta) / 2$	β	0

Los métodos anteriores empleados para producir los grupos pueden ser considerados de tres tipos. Los basados en la a) vinculación entre grupos, b) métodos de varianza y los c) métodos basados en los centroides. Además, recordemos la existencia de los métodos de carácter distribucional y los basados en la densidad. Como hemos considerado, estos métodos se diferencian en la forma como determinan la distancia entre los grupos existentes conforme avanza el proceso de aglomeración. Vamos seguidamente a describir las características de los métodos anteriores.

Single linkage o Vecino más próximo ("nearest neighbour"). Uno de los procedimientos más básicos es la agrupación según el vecino más próximo. Los primeros casos que se combinan son aquellos que tienen una distancia menor (o una proximidad mayor) entre ellos. A partir de ahí, las distancias de los otros casos hasta ese grupo se determina a partir de la distancia con el caso más próximo que ya pertenece a ese grupo. La distancia entre los casos que no han sido agrupados no varía, de forma que la *distancia entre dos conglomerados o grupos es la distancia entre los dos casos más próximos, perteneciendo cada uno de ellos un grupo distinto*. Tal y como puede observarse, en el método de vinculación simple, la distancia entre dos clústeres es la distancia mínima considerando todos los pares de casos entre los dos clústeres. Una vez estimada la distancia más próxima, se fusionan los dos grupos. El método es poco robusto, por lo que pueden influir notablemente los casos extremos. Tanto produciendo nuevos clústeres adicionales, como provocando que otros clústeres se fusionen. Es lo que se denomina como efecto de "encadenado" (chaining). Dado que los casos más próximos a cada uno de los dos grupos son los que dirigen la fusión, los clústeres resultantes pueden ser alargados y delgados. En el caso en que esta propiedad sea indeseable para el investigador, es posible recurrir a otros métodos como son complete linkage o average linkage.

Complete linkage o Vecino más lejano ("furthest neighbour"), También es posible emplear criterios alternativos (siguiendo una lógica parecida) para definir la distancia entre grupos, como es por ejemplo la técnica conocida como el *vecino más alejado*. En este método la *distancia entre los grupos es la que determine los dos casos más diferentes o distantes, perteneciendo cada caso un grupo diferente*. Este procedimiento produce el efecto contrario sobre los clústeres. Genera clústeres muy compactos espacialmente. Este efecto puede ser inapropiado si el objetivo es detectar clústeres alargados y delgados. Estos dos efectos contrarios de los dos métodos en la formación de los grupos son analizados en detalle por Kaufman y Rousseeuw (1990).

Average linkage. Otras técnicas, como el método de agrupación según la media entre grupos (UPGMA), considera la *distancia entre dos grupos como*

la media de las distancias entre todos los pares de casos en los que cada uno de ellos procede de un grupo (clúster) diferente. Este procedimiento emplea la información de todos los pares de distancias, y no solamente los de aquellos pares que se encuentran particularmente próximos o alejados. En ese sentido, es un procedimiento que incorpora mucha más información para ir definiendo la pertenencia a grupos, por lo que acostumbra ser preferido respecto a las técnicas que solamente tiene en cuenta los pares de casos más extremos, ya sea por su proximidad o lejanía.

Al igual que el método de agrupación anterior considera las distancias entre los pares definidos por los casos que pertenecen a grupos distintos, existe otra opción alternativa por la cual se combinan los grupos de forma que las distancias medias entre todos los pares de casos que pertenecerían a ese nuevo grupo se minimiza ("*Average linkage within groups method"*). Es decir, que la *distancia entre dos grupos es la media de todas las distancias entre los pares de casos posibles que formarían el nuevo grupo*. Kaufman y Rousseeuw (1990), proponen que el método de average linkage como uno de los más robustos y posiblemente el más apropiado para la mayoría de las ocasiones.

El método *Weighted average linkage* es una variación del average linkage**.** La idea básica (al igual que sucederá con median linkage) es responder a cómo se deben tratar los grupos con un tamaño desigual cuando se fusionan. En average linkage, el número de casos en cada grupo se tiene en cuenta al producir el grupo resultante de la fusión, por lo que los grupos más grandes tienen un peso mucho mayor. Este método da un peso igual a cada caso, independientemente del clúster al que pertenezca. Como su nombre indica, en weighted average, los dos grupos reciben el mismo peso para definir el grupo que resulta de la fusión, independientemente del número de casos de cada grupo. Para ello, los casos que proceden de grupos más pequeños reciben un peso mayor que aquellos casos que forman los grupos más grandes.

Centroid linkage o Agrupación de centroides ("centroid clustering"). El *método de centroides* determina la distancia entre dos grupos como la distancia entre sus medias. Es decir, este método fusiona aquellos grupos cuyas medias están más próximas. Para ello, considera las medias como una especie de centro de gravedad del grupo. Su diferencia con el método de average linkage es que, como ya hemos considerado, este último considera la distancia media entre los casos que pertenecen a los dos grupos, mientras que el método de centroide considera la distancia entre las medias de los dos grupos.

Una de las desventajas de este método es que la distancia en la que cada grupo se combina puede disminuir de un paso para el siguiente. Es decir, que los grupos que se fusionan en una etapa más avanzada son más diferentes que aquellos que fueron fusionados en etapas anteriores. Esto es una propiedad

indeseable en la medida que existen diferentes probabilidades de ser agrupados en función al momento en que se encuentra el proceso de agrupación. En este método, el centroide de un nuevo grupo que es producto de la fusión de otros grupos, se calcula como una combinación ponderada de los centroides de los dos grupos que han sido fusionados, y donde los pesos son proporcionales al tamaño de los grupos. En ese sentido, el tamaño de los grupos que son fusionados influye sensiblemente en el nuevo centroide que define el nuevo grupo. Esto es consecuencia de que todos los casos tienen un peso igual.

Median linkage. Es una variación del método de centroides. Este efecto del tamaño de los grupos en la formación de los nuevos grupos, que se produce cuando se emplea el método del centroide, puede corregirse mediante la aplicación del método basado en la mediana. En este método, los centroides de los dos grupos que son combinados, se ponderan con igual peso para calcular el nuevo centroide, independientemente del tamaño previo de cada grupo fusionado. Esto permite que los grupos pequeños tengan una mayor presencia (y peso) en el momento de caracterizar (es decir, determinar el nuevo centroide) del nuevo grupo en el que se incorporan. Esta posibilidad de caracterizar el grupo resultante de la fusión según la heterogeneidad de los grupos que se fusionan, y no según el tamaño de los grupos fusionados, es un elemento importante que debe ser decido por el investigador de acuerdo con los objetivos de la investigación.

Ward's method. Otro método usado con frecuencia es el propuesto por Ward. Aplicando el *Método de Ward,* para cada grupo se calcula la media de todas las variables. Posteriormente, y para cada caso, se calcula la distancia euclídea al cuadrado a la media del grupo. Esa distancia se suma para todos los casos. *En cada paso se agrupan los dos clúster que producen un menor incremento en la suma total de cuadrados de las distancias en los conglomerados*. En definitiva, fusiona los dos grupos que producen el incremento menor en la suma de cuadrados del error. Su enfoque es fusionar aquellos grupos que optimizan una función definida en términos objetivos. Kaufman y Rousseeuw (1990) afirman que este método funciona correctamente cuando los grupos son esféricos y multivariados normales, pero es problemático si los grupos son de diferentes tamaños o contiene un número desigual de casos.

Flexibeta. Flexible beta emplea una distancia media ponderada entre dos casos incorporados en dos clústeres diferentes para decidir los alejados que se encuentran. El investigador decide el valor de la ponderación a utilizar, dentro de un rango de -1 a 1.

*K-nbd***.** Es un método de vinculación mediante el empleo de la densidad. La densidad estimada es proporcional al número de casos en la esfera de menor tamaño que contenga el vecino más próximo al rango k. Partiendo de la estimación de densidad se construye una nueva matriz de disimilaridad. A dicha matriz de disimilaridad se aplica el método de linkage simple. El valor de k lo facilita el investigador, oscilando entre 1 y el número total de casos.

Uniform Kernel. Como ya se comentó, es un método basado en la densidad. La densidad estimada es proporcional al número de casos incluidos en una esfera de radio r. Partiendo de dicha estimación de densidad se construye una nueva matriz de disimilaridad a la que se le aplica el método de linkage simple.

No obstante, debemos recordar que todos los criterios empleados para combinar los casos en un grupo se basan en la matriz de distancia o proximidad entre ellos. A partir de esas distancias o proximidades se establecen las agrupaciones. Por ello, un mismo método para agrupar puede producir resultados diferentes dependiendo de la medida utilizada para estimar las distancias o proximidades.

En el proceso de decidir el número de clústeres el investigador puede recurrir a diferentes indicadores y mediciones que le pueden ayudar en la determinación de cuantos grupos pueden ser los más adecuados. Vamos a presentar varias mediciones que aportan información sobre el procedimiento de agrupado y las soluciones que va ofreciendo.

3.2.1. *Procedimientos de validación de los clústeres*

"Ódiame por piedad yo te lo pido
Ódiame sin medida ni clemencia
Odio quiero más que indiferencia
Pues que el odio hiere menos que el olvido.

Yo quedaré si me odias convencido
De que me amaste ayer con insistencia
Porque es cierto de que en la existencia
Tan solo se odia lo que se ha querido.

Qué vales tú más que yo hembra orgullosa?
Vales lo que tu carne blanca y dura
Pues al cabo, en el seno de la fosa,
Llevaremos la misma vestidura.

Yo que rompí la lid de la victoria
Premio de vencedores obtenido
Después de un "de" que indica pertenencia
A lo sumo llevarás otro apellido.

El odio (1908)[8]

Existen muchos coeficientes que permiten evaluar la mayor o menor validez de la solución (número de grupos que observamos en el análisis). Todos ellos comparten el mismo objetivo: identificar conjuntos de clústeres que sean compactos, con una varianza mínima entre los casos que forman el clúster, que estén bien separados entre ellos, por lo que sus medias deben de estar lo suficientemente separadas (en comparación con la varianza interna de los clústeres). Tal y como describen Milligan y Cooper (1985), son varios los índices que nos informan de la validez que presentan las diferentes soluciones que ofrezca un análisis de clústeres. En ese sentido, ayudan a determinar cuál puede ser el número óptimo de grupos presentes en los datos. Además de estos índices, es posible emplear un análisis de varianza de un factor que nos informe del estado de la heterogeneidad y la homogeneidad de los casos que se agrupan. En ese sentido, una herramienta muy importante es el análisis de varianza al que dedicaremos un apartado específico, dada la elevada relevancia de este procedimiento que aparece como complemento informativo en muchos otros análisis. Veamos en primer lugar el análisis de varianza para posteriormente presentar varios índices diseñados para intentar validar el número de clústeres extraídos.

3.2.1.1. Análisis de la varianza de un factor

Para poder explicar la función y utilidad que el análisis de la varianza tiene, especialmente en el diagnóstico de la validez de las soluciones de clústeres que se van evaluando, debemos efectuar una breve introducción.

En estadística inferencial, la media de una muestra es una aproximación al valor que esa media pueda tener en la población (μ). Es bastante improbable que la media de la población coincida exactamente con la media que estima la muestra, pero tampoco debería ser demasiado diferente. De hecho, es posible establecer un rango de valores que, de acuerdo a una probabilidad

8. Fuente: Avilés R., Carlos A. *Colección de canciones antiguas, copiadas por Carlos A. Avilés R., comenzada en Balao en las vacaciones de 1945 y terminado en Puná el domingo 24 de febrero de 1946* [manuscrito-fotocopia]. Balao, Santa Elena, 1945-1946. en Fidel Pablo Guerrero "Transferencias musicales", http://soymusicaecuador.blogspot.com.es/2011/10/el-odio-trasferencias-musicales.html

elegida, contenga entre ellos el valor de la media de la población. Ese rango de valores se denomina *intervalo de confianza*, y en esencia viene a afirmar que si de una población se extrajeran varias muestras en condiciones semejantes, y establecemos un nivel de confianza del 95%, el 95% de las muestras producirían un intervalo de confianza que incluiría el valor de la media en la población. Dado que el parámetro (μ) es desconocido, no es posible determinar si un intervalo en concreto es el que lo contiene o no.

Planteemos ahora que en la población existen grupos o clústeres para los que pensamos que las medias pueden ser diferentes. Por ejemplo, comparando las medias de ingresos de hombres y mujeres, o comparando según raza, o comparando el gasto medio según tipos de hogares o estilos de vida, etc. En el caso que se aprecie que el valor de la media es diferente entre los grupos considerados, la cuestión es si dichas diferencias pueden haberse producido por casualidad o tenemos base empírica para plantearnos que puedan ser realmente diferentes en la población. El procedimiento estadístico más usual para testar si las medias de diferentes grupos son o no iguales es el análisis de la varianza.

Como es habitual en los procedimientos estadísticos, se presupone que las medias proceden de poblaciones (grupos) independientes, con variables que muestran una distribución normal y una varianza semejante en todos los grupos. Son varios los test disponibles para examinar la homogeneidad de la varianza (Cochrans C, Bartlett-Box F, etc.) al igual que la normalidad. En todo caso, cuando los tamaños muestrales de los grupos son iguales o muy parecidos, la mayoría de los test son bastante robustos a la violación de la homogeneidad de la varianza. En el caso de no homogeneidad de la varianza o no normalidad cabría plantearse la posibilidad de transformar las variables. Una segunda opción es emplear un procedimiento no paramétrico para testar la igualdad de las medias, como puede ser el análisis de la varianza de un factor de Kruskal-Wallis.

Otra cuestión relevante es la presunción de que los grupos que comparamos son todos los grupos con interés para el investigador. Es decir, que los grupos formados (ya sea por género, raza, tipos de hogares, etc.) contienen todas las categorías que importan y no son realmente una muestra de los grupos existentes. Esta característica (que todos los grupos de interés estén considerados en la comparación) hace que se denomine modelo de efectos fijos (fixed-effects model).

En el análisis de la varianza (que ya sabemos tiene por finalidad testar la igualdad de las medias) toda la variabilidad que se observa en la variable se descompone en dos partes. Por un lado, la variabilidad interna dentro de cada uno de los grupos (por ejemplo, la variabilidad de la variable ingresos alrede-

dor de la media en el grupo de hombres, y la variabilidad de la variable ingresos en torno a su media de ingresos en el grupos de mujeres). Esta variabilidad intragrupos se mide mediante la denominada suma de cuadrados intragrupos (*within-groups sum of squares* o SSW). La idea es determinar cuanta variación interna respecto a la media existe en los diferentes grupos. Siendo K el número de grupos considerados.

$$SSW = \Sigma (N_i - 1) S_i^2$$

Siendo S_i^2 la varianza del grupo i entorno a su media y N_i el número de casos en el grupo i.

La otra variabilidad considerada es la variabilidad existente en las medias de los diferentes grupos. Esta variabilidad se mide mediante la suma de cuadrados entre-grupos o (*between-groups sun of squares* o SSB).

$$SSB = \Sigma N_i (\ddot{X}_i - \ddot{X})^2$$

Siendo N_i el número de casos en el grupo i, $\ddot{X}_i$ la media del grupo i y $\ddot{X}$ la media para el conjunto de la muestra.

Otros coeficientes que se calculan son las medias cuadráticas, que se obtienen dividiendo la suma de cuadrados por los grados de libertad. Los grados de libertad en el caso de la suma de cuadrados entre-grupos es k – 1 (siendo k el número de grupos). Por ello, la media cuadrática entre-grupos es igual a

$$\text{Media cuadrática entre-grupos} = SSB / k - 1$$

Para la suma de cuadrados intra-grupo, sus grados de libertad son el número de casos de la muestra menos el número de grupos *k*, es decir N - *k*.

$$\text{Media cuadrática intra-grupos} = SSW / N - k$$

Todos estos datos se muestran habitualmente en los resultados de un análisis de varianza. Para testar la posible igualdad entre las medias de los grupos considerados se calcula el estadístico F. Este estadístico es simplemente la media cuadrática entre-grupos dividida por la media cuadrática intra-grupos.

$$F = \text{media cuadrática entre-grupos} / \text{media cuadrática intra-grupos}$$

Para determinar si las diferencias entre medias son significativas se compara el valor F obtenido con la distribución F, para k-1 y N-k grados de libertad. El nivel de significación que se observa se refiere a la probabilidad de obtener un valor F de ese valor cuando todas las medias sean iguales en la población. Si la probabilidad es lo bastante pequeña, la hipótesis que afirma que todas las medias son iguales en la población puede ser rechazada. Por lo

general, las probabilidades de referencia son las inferiores a 0.05 o 0.01 dependiendo de la significación elegida. Cuando la significación es inferior a dichos valores, se puede rechazar la hipótesis de que las medias sean iguales en la población de la que se ha extraído la muestra.

Una utilidad especial en el análisis de clústeres es la posibilidad de efectuar múltiples comparaciones entre pares de grupos para concretar cuáles son más probablemente diferentes en sus medias. En el caso que un análisis produzca varios clústeres, hay tres cuestiones importantes que podemos responder desde el análisis de la varianza. Primero, hasta qué punto puede concluirse que estos son diferentes respecto a las variables que les han dado "forma". Es decir, que variables han tenido más peso para definir los grupos. Segundo, es interesante que los clústeres definidos sean diferentes respecto a otras variables relevantes y tercero, si entre ellos son los bastante heterogéneos (e internamente homogéneos). Las tres cuestiones que surgen del análisis de clústeres encuentran respuesta en el análisis de varianza de un factor. El factor, es evidentemente, los clústeres definidos por el análisis.

El test de comparación de medias basado en la distribución F nos indica si las medias son diferentes o no en términos estadísticos. Sin embargo, no informa si las medias de todos los grupos son diferentes entre sí, o solamente lo son las diferencias entre algunos grupos. Para esta tarea se desarrolla lo que se denomina "comparación múltiple".

Una estrategia podría ser efectuar varios contrastes mediante la aplicación de múltiples t-test para cada par de medias comparadas. Esta estrategia sin embargo, produciría errores dado que al efectuar un número elevado de contrastes, alguno de ellos aparecería como significativo a consecuencia del elevado número de pruebas, incluso siendo iguales en la población (Snedecor, 1967). Para corregir este sesgo, los procedimientos basados en comparaciones múltiples son bastante más exigentes para dar por válida una diferencia entre dos medias.

Las opciones para efectuar un análisis de múltiples comparaciones es muy elevado, correspondiendo con los diferentes enfoques utilizados para proteger contra el error de dar por significativa una diferencia que no lo es (Winer, 1971). En este caso, recomendamos el test de *Scheffé* por dos motivos: es un test especialmente conservador que exige diferencias muy importantes para dar por desiguales a las medias comparadas, y la presentación de los resultados en forma matricial, que permite una interpretación rápida y comprensiva. En la matriz se agrupan las medias que no son significativamente diferentes y se indica con cuales otras sí lo son. Algunos grupos pueden tener una media muy diferente en relación con otros grupos, o en algunos casos, tener media diferente con unos grupos sí y con otros no.

Esta presentación matricial se muestra cuando la comparación es entre tres grupos o más. En el caso de que solamente existan dos grupos, el análisis de varianza (en definitiva un t- test), ya es bastante conclusivo por él mismo. También es frecuente que el resultado del análisis muestre la fórmula que expresa como de grande debe ser una diferencia entre medias para ser aceptada como significativa en la comparación múltiple.

Tanto el análisis de varianza (paramétrico o no) como los test de comparación múltiple son de gran utilidad para estudiar el significado y potencialidades de los grupos extraídos de un análisis de clústeres.

3.2.1.2. Índices de validación de clústeres

En este texto vamos a considerar varios índices. El primero es la RMSSTD (*Root-mean-square standard deviation*). Este índice es válido cuando partimos de una matriz de datos rectangular. Cuando se utiliza directamente una matriz de disimilaridad (recordemos que es triangular), este índice solamente es válido cuando se emplean los métodos de centroide, average o Ward. Este coeficiente se calcula a partir de una combinación de las desviaciones estándar de las variables que se emplean en la definición del clúster. De acuerdo a Sharma (1995), se calcula obteniendo la suma de cuadrados intra-grupo (within-group sum of squares) de cada clúster, y después se normaliza multiplicándolo por el número de casos en el clúster así como por el número de variables.

$$\text{RMSSTD} = \sqrt{W_k} / \sqrt{(v(n_k - 1))}$$

Donde W_k nota la suma de cuadrados intra-grupo del clúster $_k$, n_k es el número de casos en el clúster k, siendo v el número de variables empleadas en el análisis de clústeres. Dado que el objetivo de un análisis de clústeres es formar grupos homogéneos, este coeficiente debería ser lo más pequeño posible. Por ello, si conforme avanza el proceso de conglomeración jerárquica el valor de RMSSTD se incrementa, indica que la nueva solución es peor que la anterior.

El índice de Dunn (1973) es otra alternativa para evaluar la validez del número de clústeres. Bezdek y Pal (1998) desarrollarían posteriormente una generalización de este enfoque. Originalmente, la distancia entre dos grupos se define como la distancia mínima entre dos casos pertenecientes a grupos diferentes, mientras que el diámetro de un grupo o clúster se define como la distancia máxima entre dos casos pertenecientes al mismo clúster. Dunn propone la siguiente medición. Consideremos que en un momento determinado la conglomeración jerárquica propone un número *k* de clústeres. Para cual-

quier par de clústeres x_i y x_j sea $\delta(x_i,x_j)$ la distancia entre los dos clústeres. Consideremos $\Delta(x_i)$ el diámetro del clúster x_i. El índice de Dunn se define como el valor mínimo de la razón entre la medida de disimilaridad de los dos clústeres y el diámetro del clúster. El mínimo se determina para todos los clústeres presentes en los datos. Este planteamiento presenta un problema específico. En el caso que uno de los clústeres este bastante disperso, mientras que el otro sea muy compacto, dado que el denominador emplea el valor máximo en lugar de algún promedio, puede producir que el valor del índice de Dunn para ese par de clústeres sea inusualmente bajo. Es algo a considerar durante el análisis. Por lo general, cuanto mayor es el valor del índice de Dunn más compacta y diferenciada es la solución que ofrece el análisis de clústeres (número de clústeres). Cuando los datos contienen clústeres muy compactos y bien separados entre ellos, la expectativa es que la distancia entre los clústeres sea elevada y el diámetro de los clústeres muy reducido. Basándonos en esa definición, valores elevados del índice corresponden con clústeres compactos y bien separados. Cuando se muestra en formato gráfico la relación entre el número de clústeres y el índice de Dunn, la solución que muestre el mayor valor en el índice debería ser la más correcta. En este caso, su empleo es adecuado tanto para matrices de datos rectangulares como de disimilaridad.

Otra estrategia para medir la validez de la solución que ofrece el número de clústeres, es el índice de Davies-Bouldin (1979). Este índice busca cuantificar la similaridad media entre un clúster y aquellos otros clústeres que puedan estar próximos a él. Sea k el número de grupos en un determinado momento del análisis jerárquico y donde Vx_i indica el centro del clúster X_i y $| X_i |$ el tamaño del clúster X_i.

Si la medición de la dispersión del clúster X_i la definimos como

$$S_i = (1/ \mid X_i \mid \Sigma d^2 (x, Vx_i))^{1/2}$$

para todo x perteneciente a X_i , y la disimilaridad entre dos clústeres (X_i y X_j) como

$$d_{ij} = d(Vx_i, Vx_j)$$

y sea $R_i = Max_{j, j\neq 1} (S_i + S_j / d_{ij})$

Entonces el índice de Davies y Bouldin es igual a $1/k (\Sigma R_i)$

De acuerdo a la formulación del índice Davies-Bouldin, cuanto menor es su valor mejor es la solución. Es decir, el número de clústeres puede ser el más apropiado. Este índice puede calcularse para matrices rectangulares de datos.

El índice *pseudo F* (Calinski y Harabasz, 1974) muestra la razón entre la varianza entre-grupos con la varianza intra-grupos. Siendo n el número de

casos y K el número de clústeres en cualquier fase de proceso jerárquico de conglomeración. Sea GSS la suma de cuadrados entre-grupos y WSS la suma de cuadrados intra-grupo, entonces

$$\text{Pseudo F} = ((\text{GSS})/ (\text{K-1})) / ((\text{WSS}) / (\text{N-K}))$$

De acuerdo a este índice, valores elevados de *Pseudo F* expresarían unos conglomerados compactos y bien diferenciados entre ellos. Los "picos" en los valores de este índice expresarían, especialmente, una gran separación entre grupos. Al igual que en el caso anterior, se acostumbra a graficar cada solución (número de clústeres) con su valor de *pseudo F*, para poder evaluar que número de clústeres puede ser el más indicado.

Este índice puede utilizarse con cualquier método jerárquico cuando se trata datos rectangulares. Cuando se utilizan matrices de disimilaridad, solamente puede emplearse este índice cuando se emplean los métodos de average, centroid y Ward.

Al igual que se emplea un *pseudo F*, es factible emplear un *pseudo t-cuadrado*. El índice basado en el *pseudo t-cuadrado* para evaluar el clúster resultado de una fusión de otros clústeres. Consideremos dos clústeres K y J que son fusionados para formar un nuevo clúster. El *pseudo t-cuadrado* vendría definido por

$$\textit{pseudo t-cuadrado} = B_{KJ} / ((W_k + W_J) / (n_k + n_j - 2))$$

siendo n_k y n_j el número de casos en el clúster K y J, W_K y W_J son las sumas de cuadrados intra-grupos de los clústeres K y J. B_{KJ} nota la suma de cuadrados entre-grupos. Este índice, como expresa su procedimiento de cálculo, mide la diferencia entre dos clústeres que han sido fusionados en un determinado momento del proceso de conglomeración jerárquica. En ese sentido, si la *pseudo t-cuadrado* cambia fuertemente en la fase t del procedimiento de conglomeración, significa que la solución de clúster en la fase t+1 es óptima.

El SPRSQ (semiparcial R-cuadrado) es una medida empleada para medir la homogeneidad resultante de fusionar dos clústeres. En ese sentido, expresa la pérdida de homogeneidad que se produce al combinar dos clústeres. Cuando los valores son bajos, indica que los grupos fusionados eran bastante homogéneos entre sí. Con la intención de medir esa posible homogeneidad de los grupos que se fusionan puede emplearse la "distancia entre los centroides". Ya sabemos que la distancia entre los centroides es simplemente la distancia euclidiana entre los centroides de los dos grupos que se estudia fusionar. En tanto que medida de homogeneidad, esta distancia debería ser baja cuando se desea que los grupos fusionados sean homogéneos entre sí.

También es posible emplear el RSQ (R cuadrado) para evaluar cómo son de diferentes dos grupos entre sí. En el caso de existir solamente un grupo, el r cuadrado será igual a cero. Por ello, valores elevados de r cuadrado expresan que dos grupos son bastante diferentes entre ellos.

Índices de validación	Matriz distancias	Matriz rectangular	Interpretación
RMSSTD	Solamente centroide, average, Ward	SI	este coeficiente debería ser lo más **pequeño** posible
Dunn	SI	SI	valores **elevados** del índice corresponden con clústeres compactos y bien separados
Davies-Bouldin (DB)	NO	SI	cuanto **menor** es su valor mejor es la solución
pseudo F (CHF)	Solamente centroide, average, Ward	SI	valores **elevados** de *Pseudo F* expresarían unos conglomerados compactos y bien diferenciados entre ellos
pseudo t-cuadrado (PTS)			Si cambia fuertemente en la fase t del procedimiento de conglomeración, significa que la solución de clúster en la fase t+1 es óptima.
SPRSQ (semiparcial R-cuadrado)			Cuando los valores son **bajos**, indica que los grupos fusionados eran bastante homogéneos entre sí
RSQ (R cuadrado)			Valores **elevados** de r cuadrado expresan que dos grupos son bastante diferentes
Silhouette coefficient			Los casos con un valor **elevado** se consideran que están bien compactados e integrados en el clúster

Silhouette coefficient. Compara la distancia media entre los elementos que forman un clúster con las distancias medias hasta los casos que forman otro clúster diferente. Los casos con un valor elevado se consideran que están bien compactados e integrados en el clúster. Los casos con valores bajos en este índice pueden ser casos extremos. Este índice funciona especialmente bien

con el método de k-medias, y es empleado para determinar el número óptimo de grupos.

Todos estos índices plantean la evaluación de la validez de modo interno, empleando los mismos datos que han sido utilizados para estimar los clústeres. Estos índices, en tanto que evaluaciones internas, son especialmente útiles para determinar si un algoritmo es mejor que otro, pero no necesariamente que produzca resultados más validos. Los procedimientos de evaluación externa, que exigen controles exógenos a la matriz de datos, no se consideran en este texto.

Una vez considerados varios de los criterios empleados más habitualmente para la fusión de grupos, y con ello creando nuevos conglomerados, es el momento de plantear algunos casos prácticos.

3.2.2. *La agrupación de casos mediante métodos jerárquicos*

I: /Ódiame por piedad yo te lo pido
Ódiame sin medida ni clemencia
Odio quiero más que indiferencia,
que el rencor hiere menos que el olvido/ (bis).

II: Yo quedaré, si me odias, convencido
De que me amaste ayer con insistencia
pues estoy cierto de que en la existencia
Tan solo se odia lo que se ha querido.

III: Qué vales más que yo niña orgullosa?
Vales lo que tu carne blanca y dura
pero al cabo, en el seno de la fosa,
Llevaremos la misma vestidura.

Más si tú en la lid de la victoria
Premio de vencedor has obtenido
Después de un "de" que indica pertenencia
A lo sumo llevarás dos apellidos.

El Odio[9]
RODOLFO MARTÍNEZ- ALFONSO DOUGARD

Veamos seguidamente un ejemplo de un proceso de aglomeración de casos. Para ello, como ya se ha considerado, partimos de la matriz de distancias que

9. Fuente: El Odio (pasillo) [disco de pizarra]/ Dúo Rodolfo Martínez- Alfonso Dougard. Disco Victor 65726-A. "Ecuatoriano Dúo con guitarra" en Fidel Pablo Guerrero "Transferencias musicales", http://soymusicaecuador.blogspot.com.es/2011/10/el-odio-trasferencias-musicales.html

hemos generado a partir de la matriz de datos original. Una diferencia importante entre ambas es que la matriz de datos original es una matriz rectangular, mientras que en el caso de las matrices de distancias o proximidades la matriz es cuadrada y simétrica. En este ejemplo exponemos una parte de la matriz de distancias, utilizando los datos presentados sobre calidad democrática en varios países de Latinoamérica.

Tabla 1. Fragmento de la matriz de distancias euclídeas al cuadrado

	Uruguay	Costa Rica	Chile	Argentina	Bolivia	Perú	Nicaragua	Ecuador
1: Uruguay	0							
2: Costa Rica	12,165	0						
3: Chile	21,33	4,583	0					
4: Argentina	47,18	14,275	6,37	0				
5: Bolivia	64,583	30,467	27,703	13,853	0			
6: Perú	37,74	14,301	5,75	4,04	15,923	0		
7: Nicaragua	62,95	40,197	37,24	27,71	5,673	21,21	0	
8: Ecuador	51,94	22,627	15,33	6,48	4,513	4,12	8,51	0
9: Brasil	61,42	26,039	19,07	6,34	3,583	6,62	10,89	0,9
10: El Salvador	27,33	18,555	29,26	33,83	22,453	24,97	18,86	21,71
11: Paraguay	55,44	31,065	28,29	19,46	3,743	13,78	1,19	4,14
12: Panamá	60,54	25,457	16,29	4,78	7,813	4,06	15,63	1,42
13: Rep. Dominicana	75,06	52,823	49,51	38,38	10,423	29,34	1,05	14,06
14: México	49,82	26,625	16,81	12,26	15,993	3,26	15,23	4,02
15: Venezuela	84,73	50,829	39,94	24,05	12,013	16,23	9,82	7,05
16: Colombia	66,09	43,207	34,8	26,85	18,003	13,75	11,08	8,93
17: Honduras	75,31	47,491	45,9	33,63	15,303	23,25	10,62	13,11
18: Guatemala	99,49	74,155	74,34	60,51	31,493	44,73	19,94	30,59

Esta es una matriz de disimilaridades

Partiendo de esta matriz de distancias, se elige el método de agrupación que se prefiera. En este ejemplo se ha elegido la vinculación media entre grupos. Como resultado de este sistema jerárquico para agrupar, se produce una paulatina formación de clústeres. Existen varios procedimientos para que el investigador pueda evaluar cómo se van formando progresivamente los grupos,

tanto en forma numérica como gráfica. Ciertamente las presentaciones gráficas son dificultosas cuando parten de la agregación desde el nivel de caso. Este es uno de los motivos por lo que el empleo de métodos jerárquicos de conglomeración son especialmente apropiados cuando el análisis no excede de unos 200 casos.

Recordemos que el análisis de conglomerados es (en el caso de métodos jerárquicos) en gran parte exploratorio. Por ello, son varias las presentaciones gráficas de la misma información de forma que ayude al investigador a decidir el número de grupos, como por ejemplo representando la forma de un árbol (Hartigan, 1975). En ese sentido, el dendrograma (Sokal and Sneath, 1963) es una expresión gráfica de este proceso de agrupación de casos y clústeres. En este ejemplo, veremos que en el lado izquierdo aparecen los países, y se aprecia cómo van incorporándose nuevos casos a grupos existentes, formando nuevos grupos o combinándose dos grupos en uno. Así, Brasil, Panamá y Ecuador forman un grupo rápidamente. Perú y México otro grupo diferente, que se unen al formado por Brasil, Ecuador y Panamá, en un paso posterior. Al grupo formado por los cinco países anteriores se une Argentina más tardíamente.

Recordemos que esta paulatina agrupación y combinación de países se produce sobre la base de la matriz distancias (según la medida elegida y tras decidir transformar o no los valores y los coeficientes), y del procedimiento escogido para determinar el cálculo de la distancia a la que se combinan los casos y los grupos. Para llegar al dendrograma el investigador ha debido tomar ya cuatro decisiones relevantes (qué variables, qué transformación, qué coeficiente de distancia, y qué método de agregación). En esta última decisión, el método de agrupación, se está decidiendo cuánta diferencia integramos en un mismo grupo[10]. En términos paradójicos, cuánta heterogeneidad se admite dentro de un grupo que pretendemos homogéneo. Al final, a la derecha del gráfico, todos los países han sido integrados en un solo grupo. Empleando la información del proceso de agregación debe decidirse cuántos grupos consideramos significativamente diferentes (es decir, que los países que los forman están próximos entre sí y diferenciados de otros grupos).

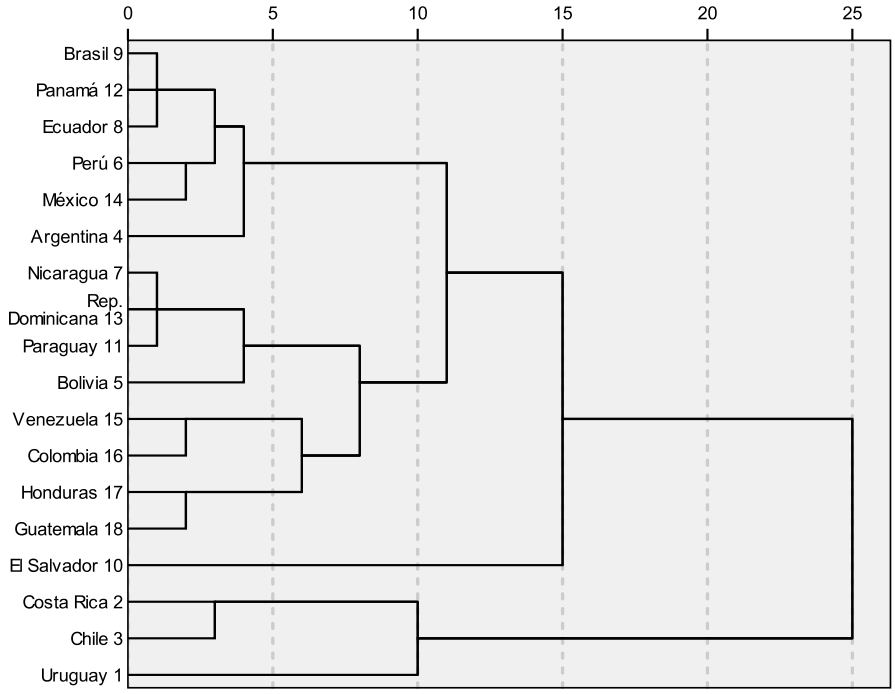
Sobre la base del dendrograma siguiente, parece observarse al menos dos soluciones diferentes. Una de estas soluciones ofrecería cuatro grupos de países, más El Salvador y posiblemente Uruguay como casos atípicos. Un

10. En alguna ocasión, el dendrograma no puede producir una combinación en la que las distancias se incrementan progresivamente. En esta situación, es posible apreciar que algunas ramas no llegan a conectarse unas con otras. En estos casos cabe plantearse optar por los métodos de vinculación simple o vinculación completa (Single o Complete linkage) según Fisher y Van Ness (1971).

grupo podría ser Brasil, Panamá, Ecuador, Perú, México, Argentina. Otro grupo Nicaragua, Rep. Dominicana, Paraguay y Bolivia. El tercer clúster puede definirse por Venezuela, Colombia, Honduras y Guatemala. El cuarto clúster, por Costa Rica y Chile.

Otra solución es decidir que son relevantes tres grupos. El grupo 1 formado por Brasil, Panamá, Ecuador, Perú, México, Argentina. El grupo 2 formado por Nicaragua, Rep. Dominicana, Paraguay, Bolivia, Venezuela, Colombia, Honduras y Guatemala. El tercer grupo por Costa Rica, Uruguay y Chile. Queda como país más atípico El Salvador.

Dendrograma que utiliza una vinculación media (entre grupos)
Combinación de conglomerados de distancia re-escalados



La decisión sobre cuántos grupos son significativos (en el sentido de que no contienen una heterogeneidad excesiva) la decide siempre el investigador. Evidentemente con posterioridad se pueden efectuar otros análisis para comprobar la significación estadística de la discriminación entre grupos. En este caso, el método más adecuado consiste en el análisis de varianza, que consideraremos más tarde.

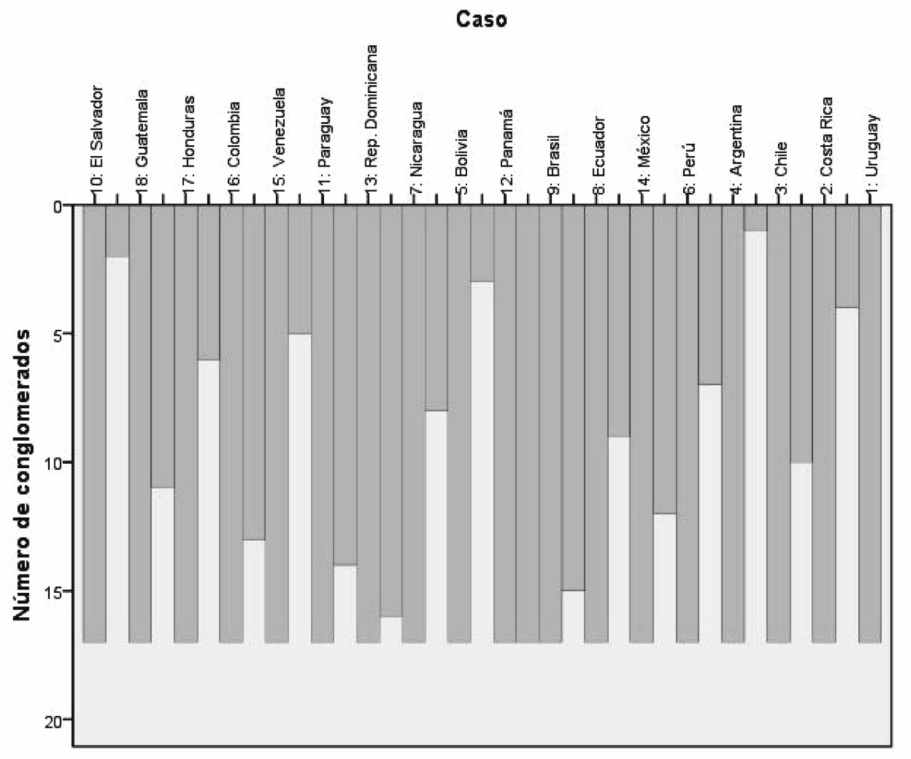
Otra información frecuente en los modelos jerárquicos es el historial de agrupación (conglomeración). En esta ocasión, los países vienen expresados por el código que los identifica. Esta información se recoge en las columnas "conglomerado que se combina". Por ejemplo, el código 9 es Brasil y el 12 Panamá.

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	9	12	,840	0	0	3
2	7	13	1,050	0	0	4
3	8	9	1,160	0	1	9
4	7	11	2,405	2	0	10
5	15	16	3,100	0	0	12
6	6	14	3,260	0	0	9
7	17	18	3,940	0	0	12
8	2	3	4,583	0	0	14
9	6	8	5,017	6	3	11
10	5	7	6,613	0	4	13
11	4	6	6,780	0	9	15
12	15	17	9,165	5	7	13
13	5	15	13,556	10	12	15
14	1	2	16,747	0	8	17
15	4	5	17,647	11	13	16
16	4	10	24,331	15	0	17
17	1	4	41,874	14	16	0

En la primera columna aparece la etapa del procedimiento. Así, en la etapa 1, se combinan los conglomerados 9 (Brasil) y 12 (Panamá). El coeficiente de .84 se refiere a la disimilitud entre los dos países. Si comprobamos en la matriz de distancias, es exactamente la distancia euclídea cuadrada entre los dos países. En el paso 2 se agrupan los dos países con menor distancia entre ellos, el 7 (Nicaragua) y el 13 (Rep. Dominicana) con un 1,05. Y así sucesivamente. En dicho cuadro puede observarse en qué orden han ido agrupándose los países en función a la medida de distancia elegida. Esta misma información se expresa en una presentación grafica denominada "Iceplot", dado que recuerdan los témpanos de hielo.

En este gráfico, en la parte superior aparecen los países, y entre ellos, de forma sombreada, cómo se agrupan. En la parte superior todos están unidos en un solo grupo, formando un clúster. Es la etapa final del análisis. Los casos, en esta ocasión países, se van agrupando desde la parte inferior. Así puede observarse como los dos países que primero aparecen unidos por dicha área sombreada son Brasil y Panamá. Los dos países que se agrupan a continuación son República Dominicana y Nicaragua.



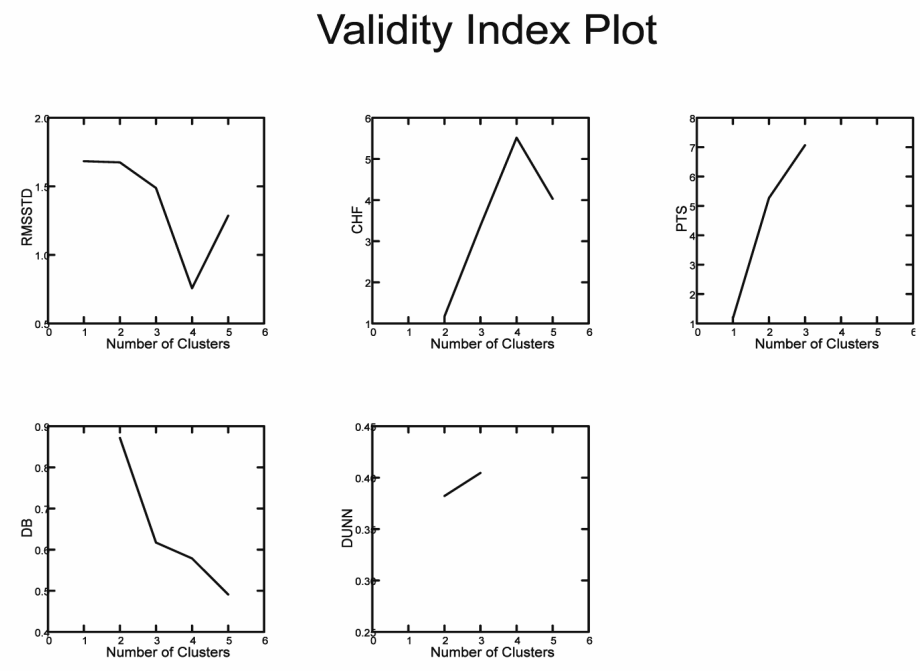
En definitiva, la información se expresa de forma diferente para ayudar al investigador en la interpretación lógica de la agrupación. En todo caso, en la medida que con posterioridad se puede evaluar la consistencia explicativa o diagnóstica de las agrupaciones, siempre puede retomarse el análisis y comprobar agrupaciones alternativas que incorporen menos heterogeneidad dentro del clúster. Cabe recordar que son varias las decisiones importantes:

a) Qué variables se emplean para determinar los grupos. Con ello decidimos el significado teórico que estos grupos pueden tener.
b) Qué medida de similitud o disimilitud se va a emplear

c) Si los datos van a ser transformados antes de calcular las similitudes o disimilitudes
d) Si los coeficientes de las medidas de similitud o disimilitud van a ser normalizados
e) Qué procedimiento de los existentes vamos a elegir para que determine la distancia entre países y grupos. Con ello decidimos como se construyen los conglomerados y la heterogeneidad que contienen.

Todas estas opciones están disponibles en la mayoría de los programas de análisis que efectúan estimación de clústeres. La otra opción que se plantea desde el inicio es la posibilidad de establecer clústeres entre variables, mediante su matriz de distancias o proximidades. Recordemos que las medidas de distancia pueden estimarse tanto para casos como para variables. La agrupación de variables nos indicará cuáles están más próximas entre sí, considerando los valores que los casos presentan en cada una de ellas.

Si tenemos en cuenta los índices de validez, RMSSTD propone cuatro clústeres, al ofrecer en esta solución el coeficiente más bajo. El índice pseudo F (CHF) señala en la misma dirección, al corresponder con cuatro clústeres su coeficiente más elevado. Pseudo t-cuadrado (PTS) cambia fuertemente del clúster dos al clúster tres, sugiriendo como solución cuatro clústeres.



Davies-Bouldin (DB) muestra un valor también bajo y consistente con los otros índices en la solución de cuatro clústeres. Dunn muestra la aproximación de un coeficiente elevado con cuatro clústeres. En conjunto, la solución de cuatro clústeres parece ser la más consistente con los datos. Vamos seguidamente a considerar el análisis de clústeres de variables.

3.2.3. *Agrupación de variables mediante métodos jerárquicos*

I. Ódiame por piedad yo te lo pido
Ódiame sin medida ni clemencia
Odio quiero más que indiferencia
que el rencor hiere menos que el olvido.

II. /Yo quedaré si me odias convencido
De que me amaste ayer con insistencia
Porque es muy cierto de que en la existencia
Tan solo se odia lo que se ha querido./. (bis)

III. /Qué vales más que yo niña orgullosa?
Vales lo que tu carne blanca y dura
Pero al fin, en el seno de la fosa,
Llevaremos la misma vestidura./ (bis)
Bis II

IV. De amores y odios [……, ilegible] solo te pido
Mi espantosa vida y compañera
Fue una pobre mujer, una cualquiera
Y a mi vida y amor partió conmigo.

V. Ahora que tengo mí triunfo asegurado
Me aconsejan que te bote de mi lado
Y una mujer así deshonra y calla/pero el triunfo no autoriza entre canallas/
(bis).

Odio[11]
SEBASTIAN ROSADO

Cuando la intención es producir una agrupación de variables según su similitud, una de las opciones es partir de su matriz de correlación. De esta forma, las variables con una correlación mayor estarían más próximas entre ellas que aquellas otras variables cuyo coeficiente de correlación sea menor o no signifi-

11. Fuente: Odio [disco de pizarra]) / Sebastián Rosado. Disco Favorite Record AKT-Ges. Linden. Precioso Record 1-45052. "Ecuador song" en Fidel Pablo Guerrero "Transferencias musicales", http://soymusicaecuador.blogspot.com.es/2011/10/el-odio-trasferencias-musicales.html

cativo. Dado que adoptamos la correlación como medida de proximidad, en muchas ocasiones el signo carece de interés. Por ello, para la matriz de similitud tomaremos el valor de las correlaciones en términos absolutos. Es una decisión tomada para este análisis en particular, pero sin embargo, en otras circunstancias puede ser más interesante conservar la información que aporta el signo de la correlación. Pensemos el caso en que el investigador tenga interés en agrupar aquellas variables que muestran una correlación positiva y diferenciarlas de otras negativas. En estas circunstancias, es importante conservar el signo dado que representa un elemento de interés teórico para el investigador.

El procedimiento de agrupación es el mismo para variables que para casos. Comienza considerando tantos grupos como variables existen y a cada paso sucesivo, las variables van formando grupos según el criterio que se haya adoptado para establecer la proximidad entre ellas.

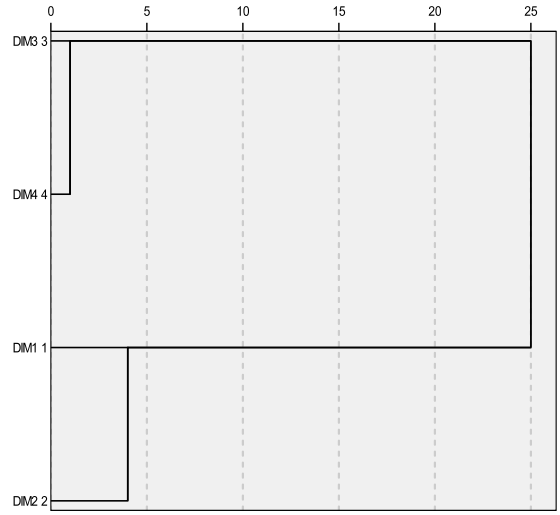
Por lo general, los resultados de la agrupación de variables que produce el análisis de conglomerados es semejante al que podríamos obtener mediante el análisis factorial, si bien con algunas diferencias. Así, en el análisis factorial existe un modelo teórico que respalda la intención de agrupar unas variables y no otras. Se presupone que los factores (agrupación de variables), en términos confirmatorios, expresan una medición concreta. En el análisis de conglomerados, la agrupación se produce de forma completamente exploratoria, en virtud de las proximidades que se aprecien entre las variables y sin necesidad de ninguna teoría previa. Una segunda diferencia importante es de carácter técnico. Así, la agrupación de variables que se produce mediante el análisis factorial, puede contener variables tanto con correlación positiva como negativa. En el caso del análisis de conglomerados, la agrupación se efectuará solamente para variables con correlación positiva. De utilizar las variables con diferentes correlaciones entre ellas (positivas y negativas), se agruparán por un lado las variables con correlación negativa y por otro las positivas. No aparecerán grupos de variables que presenten entre ellas correlaciones negativas y positivas. Por eso, cuando el signo de la correlación no es significativo, es importante efectuar el análisis de conglomerados de las variables tomando el valor absoluto de las correlaciones, con lo que las variables se agruparán en función a la similitud que existe entre ellas según determine la magnitud de su correlación. De esta forma, mediante el valor absoluto de las correlaciones, se elimina el efecto del signo.

Cabe destacar, que la agrupación de variables resultado de la administración de un análisis factorial no tiene por qué, necesariamente, coincidir con el resultado de un análisis de conglomerados. Siempre es interesante, cuando existe, una coincidencia en los resultados que generan ambos métodos. Sin embargo, esto no tiene por qué ser necesariamente así.

Matriz de distancias entre variables (Calidad democrática)				
Variables	Variables			
	DIM 1	DIM 2	DIM 3	DIM 4
DIM 1	,000			
DIM 2	78,643	,000		
DIM 3	480,409	366,012	,000	
DIM 4	22,169	18,842	5,332	,000

De acuerdo a las distancias que se determinan entre las variables, observamos como las dimensiones DIM3 y la DIM4 se agrupan rápidamente (son muy próximas), mientras que la formación de otro grupo por parte de las dimensiones DIM1 y DIM2 se hace más tardíamente. Es decir, que las dimensiones DIM1 y DIM2 se encuentran a mayor distancia entre sí que la DIM3 y la DIM4. La unificación de los dos grupos (el formado por las dimensiones DIM3 y DIM4, y el formado por las dimensiones DIM1 y DIM2) se hace al final del proceso. Con ello se expresa que existe una heterogeneidad muy elevada, o dicho en otras palabras, una relación débil entre los dos grupos de variables.

Dendrograma que utiliza una vinculación media (entre grupos)
Combinación de conglomerados de distancia re-escalados

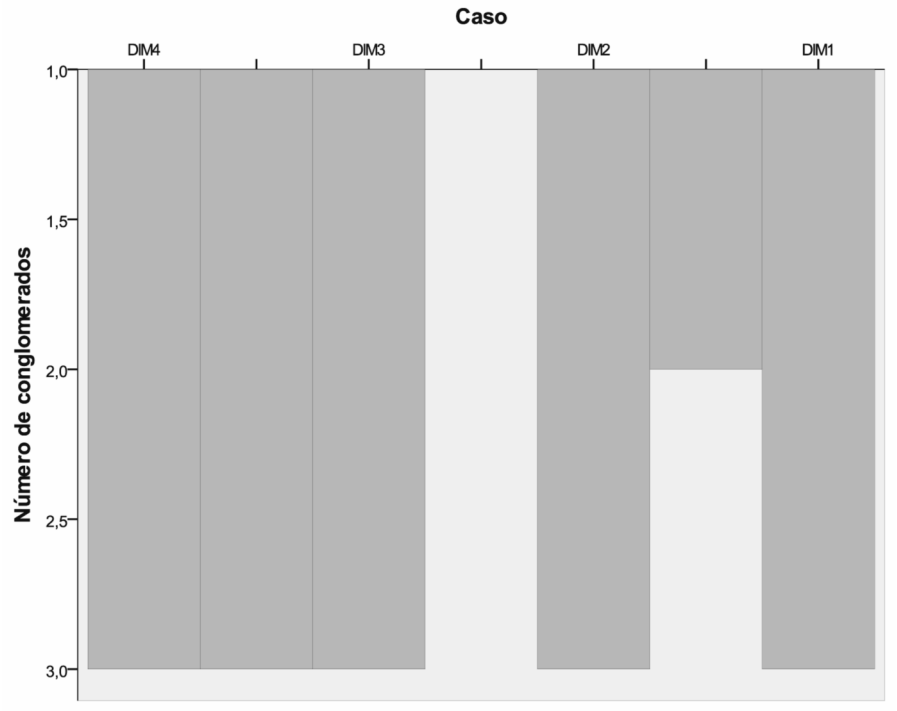


En el historial de conglomeración nuevamente podemos observar como las dos dimensiones más próximas (DIM3 y DIM4) se unen a una distancia de 28,43. Las dos dimensiones que se unen más tarde (DIM1 y DIM2) lo hacen a una distancia de 78,6. La fusión de los dos clúster se realiza mucho más tarde, a una distancia de 423,2.

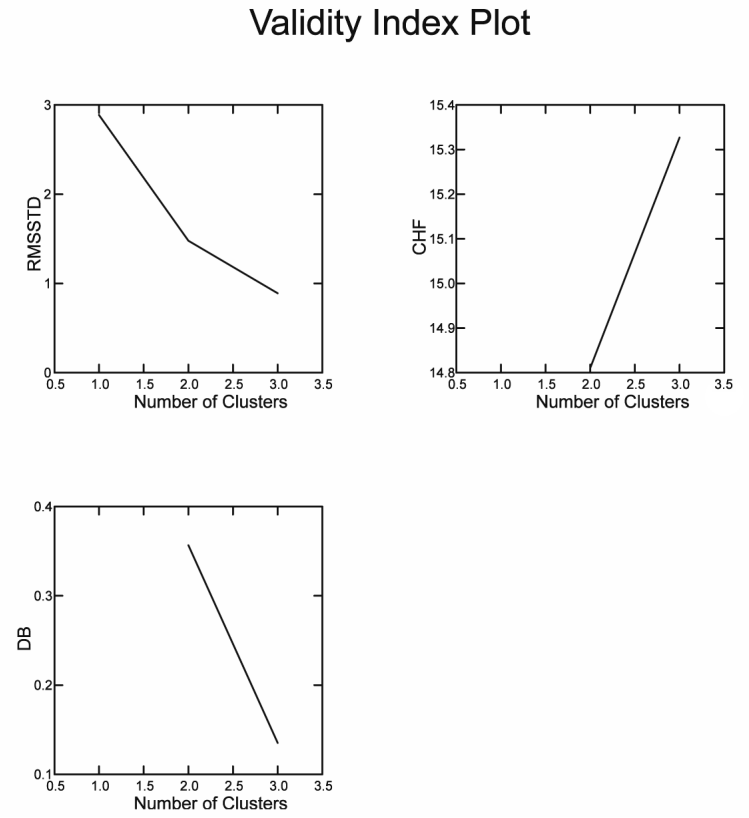
Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	3	4	28,433	0	0	3
– 2	1	2	78,643	0	0	3
3	1	3	423,230	2	1	0

Nuevamente, esa misma información viene expresada de forma gráfica en el denominado "Iceplot" o "Témpano".



En términos de validez, considerando los índices de RMSSTD, DB y la pseudo F la conclusión es que puede afirmarse que muy probablemente existen tres grupos, es decir, tres dimensiones principales diferentes. El gráfico siguiente muestra el valor de los índices mencionados al efectuar el clúster de variables.



Como hemos podido apreciar, los procedimientos de formación de conglomerados o clúster son aplicables tanto a los casos (sean individuos, países, ciudades, asociaciones, etc.) como a las variables o indicadores que se empleen para medir sus características. Este doble uso de la formación de clústeres aproxima técnicas como son el análisis factorial y el análisis de conglomerados.

3.2.4. *La conglomeración de variables y casos*

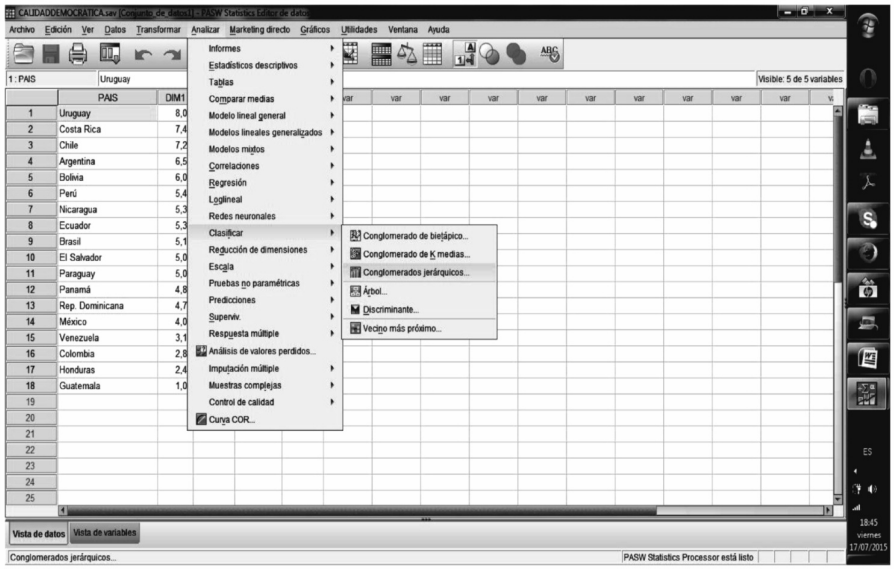
Una posibilidad es efectuar un análisis que combine la estimación de grupos de casos y de variables simultáneamente. Como ya sabemos, el análisis de

clúster es un procedimiento diseñado para detectar grupos de casos y de variables. También es posible considerar simultáneamente los casos y las variables. Es decir, la matriz de filas y columnas. Para agrupar filas y columnas simultáneamente, es preciso estandarizar primero las variables para darles a todas ellas el mismo peso. De esta forma, todas tendrán una oportunidad igual de expresar su influencia sobre los diferentes casos. Tras la estandarización, es adecuado emplear distancia euclídea con linkage simple.

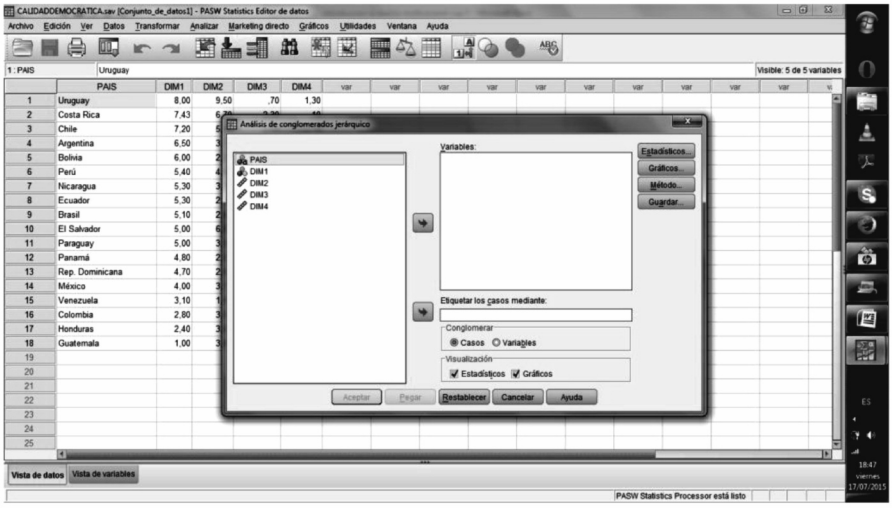
Por lo general, el resultado del análisis se puede expresar con un gráfico donde se muestra la matriz de datos, si bien permutando las filas y variables para mostrar la agrupación de casos y variables.

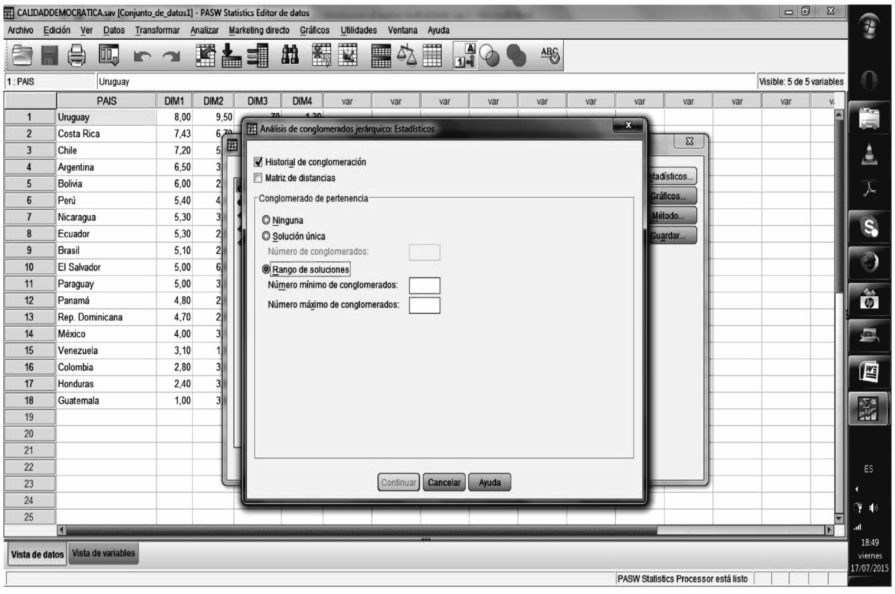
3.2.5. *Ejemplos en SPSS y SYSTAT*

La realización de estos análisis requiere del apoyo de programas informáticos. En ese sentido, son muchas las ofertas de programas tanto comerciales como no comerciales. Vamos a considerar dos programas comerciales de uso bastante extendido: SPSS y SYSTAT. Veamos seguidamente como se efectúa un análisis de conglomerados con SPSS. La opción de este análisis se encuentra en la categoría analizar, y después la opción clasificar. Esta ubicación en el menú del programa puede variar según versiones, dado que este programa ha reclasificado varias veces su sistema de menús.



En la pantalla tras elegir este análisis, puede escogerse las variables que participarán en el análisis, así como otras opciones analíticas. En este momento, la más relevante es la decisión sobre si los conglomerados se van a efectuar operando con las variables o con los casos.





Las opciones que tenemos disponibles en esta ventana se refieren al "historial de aglomeración", en el cual se recoge el paso en el que los grupos se van combinando, así como a la distancia que lo hacen, tal y como mostrábamos en el cuadro anterior. La opción "matriz de distancias" visualiza las distancias entre los objetos considerados. En el caso de ser variables mostrará una matriz cuadrada con las variables. De haber seleccionado casos, la matriz de distancias cuadrada indicará la distancia entre casos. La distancia que se calculará depende de la que haya sido seleccionada en la primera ventana.

Otra opción interesante es indicar a qué grupo pertenecería cada caso según el número de clústeres elegidos. La opción "ninguna" elimina de los resultados esta información. Al elegir una "solución única" (es decir, un número determinado de clústeres), registrará la pertenencia de cada caso a cada uno de los clústeres especificados. En este caso debe advertirse que se desea una solución con más de un clúster. Otra posibilidad es elegir un "rango de soluciones". En este caso, se estimarán varios clústeres, indicando la pertenencia a cada uno de ellos de cada caso. Los valores deben ser superiores a uno y el número mínimo de conglomerados menor (obviamente) que el número mayor. Si tomamos el ejemplo de calidad democrática, podemos apreciar como bajo la columna 2 conglomerados se indican cifras 1 y 2 mostrando la pertenencia de cada caso a cada uno de los dos conglomerados. En el otro extremo, en la columna 5 conglomerados, las cifras van desde 1 hasta 5, indicando a cuál de los cinco conglomerados pertenece cada caso.

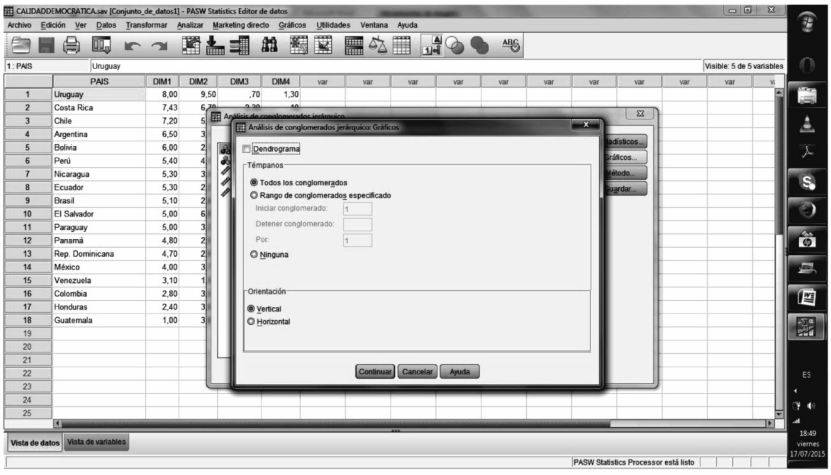
Conglomerado de pertenencia

Caso	5 conglomerados	4 conglomerados	3 conglomerados	2 conglomerados
1	1	1	1	1
2	2	1	1	1
3	2	1	1	1
4	3	2	2	2
5	4	3	2	2
6	3	2	2	2
7	4	3	2	2
8	3	2	2	2
9	3	2	2	2
10	5	4	3	2

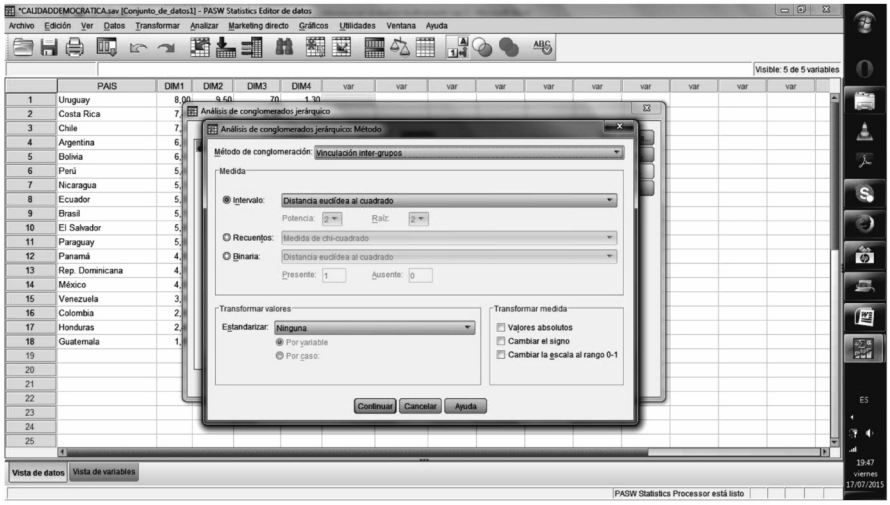
Caso	5 conglomerados	4 conglomerados	3 conglomerados	2 conglomerados
11	4	3	2	2
12	3	2	2	2
13	4	3	2	2
14	3	2	2	2
15	4	3	2	2
16	4	3	2	2
17	4	3	2	2
18	4	3	2	2

En lo que se refiere a los resultados gráficos, la opción dendrograma muestra el gráfico de agrupación visto anteriormente. En el caso del gráfico de "témpanos" es posible elegir cuántos clústeres se desea visualizar. Así la opción "todos los conglomerados" indicará el proceso de agrupación de todos los clústeres. Es posible establecer un rango de conglomerados para ser mostrados, así como el ritmo de aglomeración. Por ejemplo, indicando iniciar en 2 y terminar en 10, en saltos de 2, el gráfico mostrará la solución para 2, 4, 6, 8 y 10 clústeres. También es posible eliminar el gráfico de témpanos, o decidir la orientación vertical u horizontal.

Los gráficos son una utilidad para poder visualizar el proceso de agrupación y las distancias en las que se efectúan. Debemos recordar que cuanta más distancia, más heterogeneidad se incorpora al conglomerado.



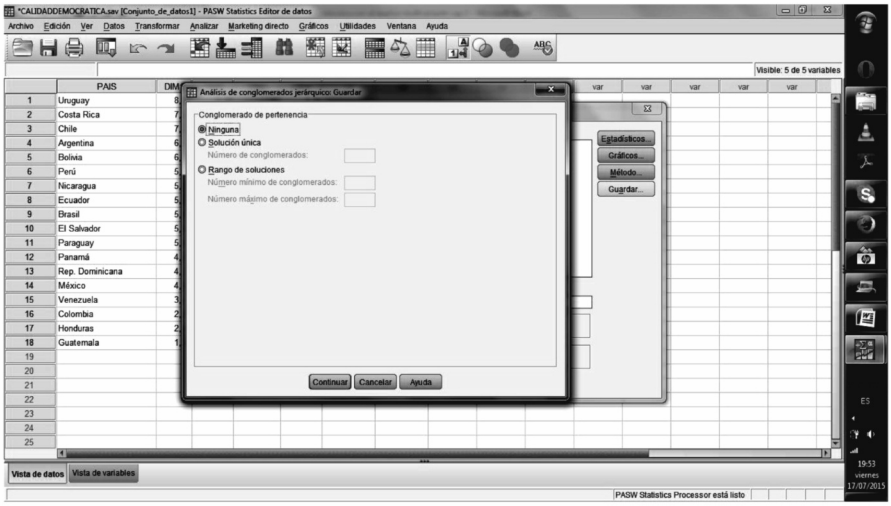
En la opción método encontramos las operaciones ya tratadas anteriormente. Así, podremos elegir el método que se prefiera de conglomeración, la distancia que se desea emplear (de intervalo, recuentos o binaria) y para cada una de ellas las diferentes medidas de proximidad o distancia. Así mismo, encontramos las opciones de estandarizar y normalizar las variables o los casos antes del análisis. Transformar las medidas recordemos que consiste en modificar los coeficientes de distancia o proximidad que han sido calculados para cada par de objetos.

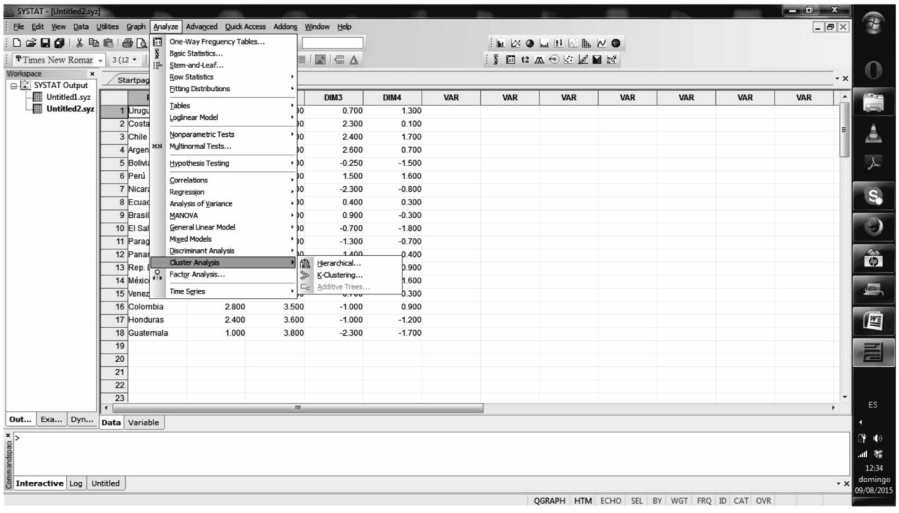


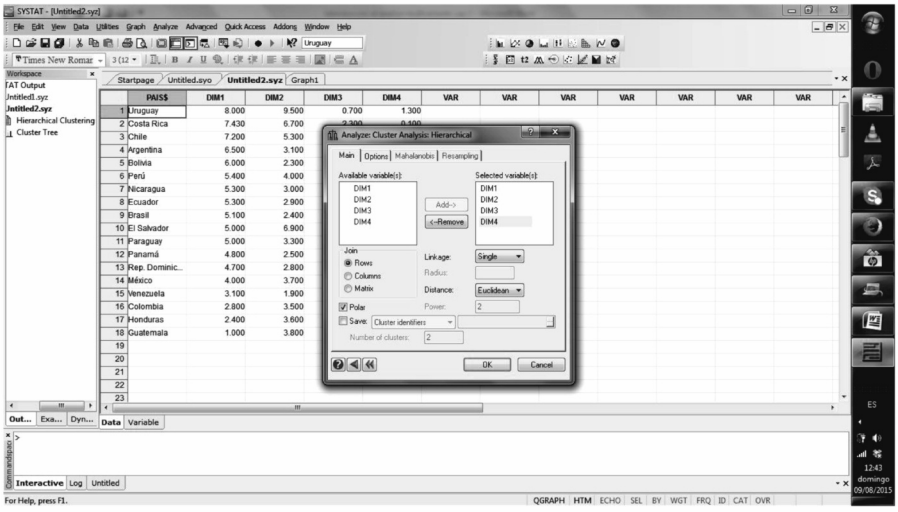
Por último, y solamente para la opción de efectuar clústeres con los casos, tenemos la opción de crear una nueva variable donde se indique la pertenencia de cada caso a cada uno de los conglomerados estimados. Nuevamente, la opción es crear una variable ("solución única"), donde se indique la pertenencia de cada caso a los clústeres decididos, o crear un conjunto de variables, donde cada una de ellas muestra la pertenencia de cada caso según el número de clústeres en esa solución ("Rango de soluciones").

En el caso de utilizar el programa SYSTAT la organización de los menús es bastante similar al programa anterior. La elección en el menú de la opción "Analizar" nos ofrece la opción "análisis de clústeres", y dentro de ella las opciones de "jerárquicos" y "no jerárquicos" (K-clústeres).

Dentro de la opción "Jerárquicos", la mayor parte de las opciones ya son conocidas. Tanto las opciones para elegir el método para estimar las distancias entre clústeres, como la elección de la distancia elegida, el efectuar el clúster de filas (rows), que indican por lo habitual los casos, o de columnas (columns) expresando variables son semejantes en los dos programas.

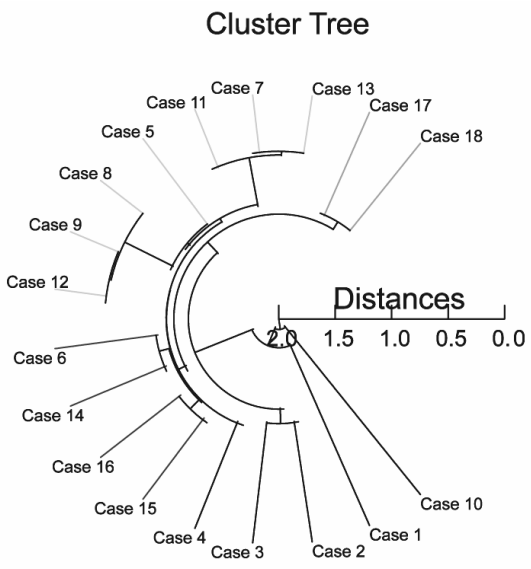

*CALIDADDEMOCRATICA.sav [Conjunto_de_datos1] - PASW Statistics Editor de datos
Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda
Visible: 5 de 5 variables
PAIS
Uruguay
Costa Rica
Chile
Argentina
Bolivia
Perú
Nicaragua
Ecuador
Brasil
El Salvador
Paraguay
Panamá
Rep. Dominicana
México
Venezuela
Colombia
Honduras
Guatemala
Análisis de conglomerados jerárquico: Guardar
Conglomerado de pertenencia
Ninguna
Solución única
Número de conglomerados:
Rango de soluciones
Número mínimo de conglomerados:
Número máximo de conglomerados:
Continuar
Cancelar
Ayuda
Estadísticos...
Gráficos...
Método...
Guardar...
Vista de datos
Vista de variables
PASW Statistics Processor está listo


SYSTAT - [Untitled2.syz]
File Edit View Data Utilities Graph Analyze Advanced Quick Access Addons Window Help
One-Way Frequency Tables...
Basic Statistics...
Stem-and-Leaf...
Row Statistics
Fitting Distributions
Tables
Loglinear Model
Nonparametric Tests
Multinormal Tests...
Hypothesis Testing
Correlations
Regression
Analysis of Variance
MANOVA
General Linear Model
Mixed Models
Discriminant Analysis
Cluster Analysis
Factor Analysis...
Time Series
Hierarchical...
K-Clustering...
Additive Trees...
DIM3
DIM4
VAR
Colombia 2.800 3.500 -1.000 0.900
Honduras 2.400 3.600 -1.000 -1.200
Guatemala 1.000 3.800 -2.300 -1.700
Data
Variable
Interactive
Log
Untitled
QGRAPH HTM ECHO SEL BY WGT FRQ ID CAT OVR

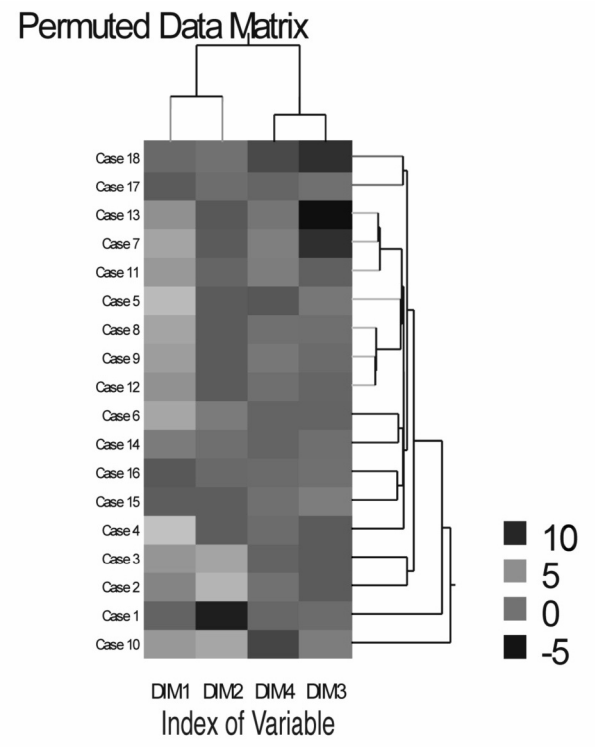


Aparecen, sin embargo, dos opciones especialmente interesantes, que implican a la expresión gráfica. El programa SYSTAT ofrece la oportunidad de mostrar la agrupación mediante un gráfico polar como el que se muestra seguidamente, para el ejemplo de la agrupación de países según calidad democrática. Es la opción "polar", que se encuentra encima de la opción "guardar" en una variable nueva el clúster de pertenencia.

Cuando no se solicita la opción "polar", se muestra el dendograma clásico.



La otra opción muy interesante es la que resulta de la elección de estimar los grupos considerando tanto los casos como las variables. Es la opción "Matriz" (tras las opciones de por fila o columna). Este procedimiento permite graficar la formación de los clústeres de variables y de casos simultáneamente, tal y como se aprecia en el gráfico siguiente.

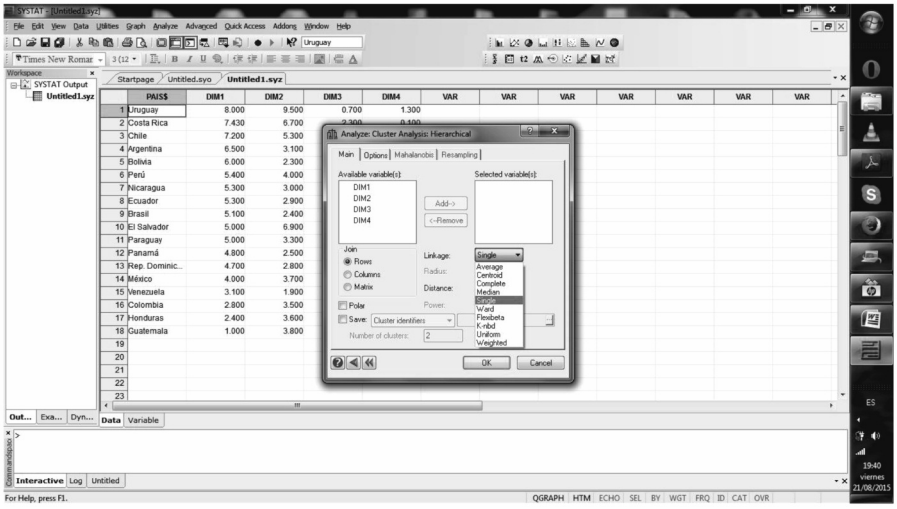


Como podemos apreciar, SYSTAT genera un gráfico donde se muestra la matriz original de datos, con las filas (casos) y columnas (variables) permutadas según el algoritmo propuesto por Gruvaeus and Wainer (1972). Los diferentes rasgos se expresan asociados a la magnitud de los valores en la matriz (Ling, 1973). La leyenda muestra los rangos de valores expresados mediante los diferentes sombreados, donde los puntos de corte entre rangos se deciden de forma que optimicen el contraste. Para decidir esos puntos de corte que optimizan, se ordenan los datos y se localizan los "saltos" más elevados entre ellos. Solamente se consideran los puntos de corte que son significativos asintóticamente al 0.05. Posteriormente se emplea el método de Tukey para determinar cuántos rangos y que características se les asocian, (Wainer and Schacht, 1978).

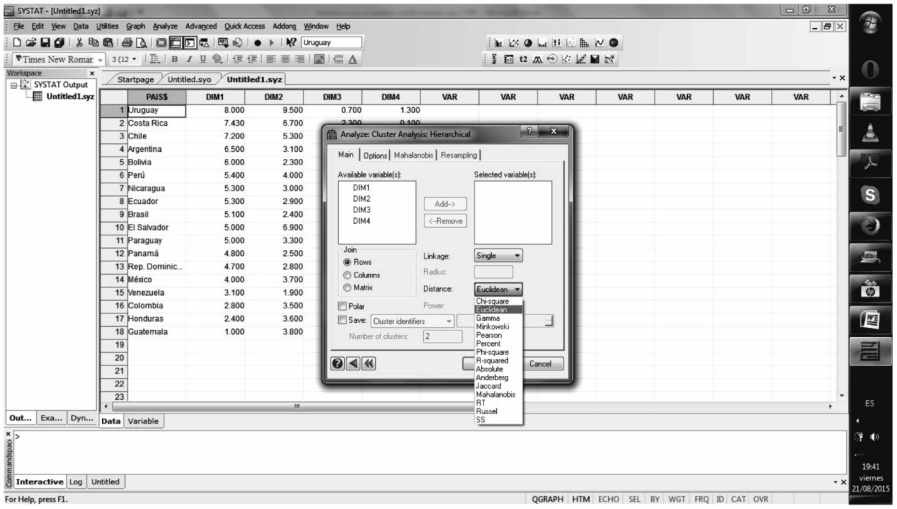
El programa SYSTAT, ofrece varias opciones para estudiar más en profundidad la formación de los clústeres. Son las que se muestran en la pestaña de "opciones". En la columna izquierda aparecen las opciones para controlar la profundidad y color del dendrograma, según varios indicadores, como son las distancias, pero también el número de casos en cada clúster. En ese sentido, es más flexible al incorporar la opción del color como elemento sustantivo para reconocer el proceso de agrupación.

En las pestañas de vinculación (linkage) y distancia (distance) se puede elegir entre varios métodos de vinculación y procedimientos de cálculo de distancias. Los métodos de vinculación (Linkage) en SYSTAT, permiten elegir entre Single, Complete, Average, Centroid, Median, Ward's (Ward, 1963), Weighted Average and Flexible Beta.

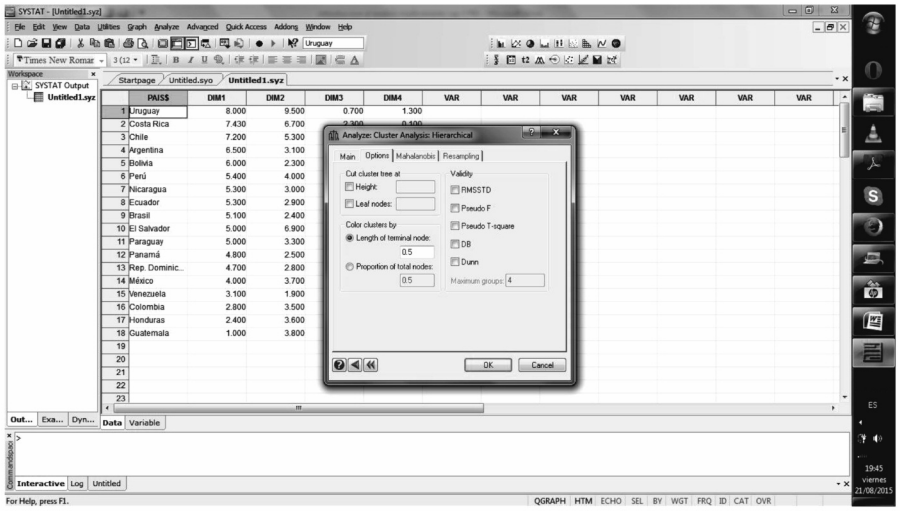
SYSTAT ofrece dos métodos para efectuar la agrupación en base a la densidad. Uno de ellos es "Uniform" y el otro es "K-*nbd*" (vecino más próximo). En ambos métodos se estima una probabilidad de densidad para los casos, y se construye una nueva matriz de disimilaridad (combinando la estimación de probabilidad y la matriz de disimilaridad original). Ambos métodos fueron explicados en páginas anteriores.



Una cuestión a tener en consideración es que las distancias no aparecen ordenadas según la métrica de las variables, como puede apreciarse en el desplegable del menú.



Por último, en lo que estamos considerando en este texto, la pestaña de "opciones" permite controlar cómo se muestra el gráfico de árbol, en el área izquierda. En el área derecha aparecen diferentes índices para medir la validez de las diferentes soluciones de clústeres.



Los índices que calcula son RMSSTD, Pseudo F, Pseudo T-square, DB o Davies-Bouldin, y por último Dunn. Por ejemplo, en la conglomeración jerárquica SYSTAT calcula el índice RMSSTD a cada paso, facilitando una

medida de la homogeneidad de los clústeres que se han formado en ese paso. Produce, asimismo, un gráfico con el valor de RMSSTD a cada paso. Con ello se puede explorar las diferentes soluciones analizando los saltos que se pueden observar en el índice. Todos ellos han sido comentados anteriormente. La última opción es el número de clústeres, por el que se indica cuántos clústeres queremos que evalúe como máximo. Por ejemplo, de elegir cinco, los índices se calcularán para un máximo de cinco clústeres.

La última pestaña que comentamos es la Mahalanobis. En dicha pestaña, se puede especificar la matriz de covarianzas para poder calcular la distancia Mahalanobis. Consideremos a continuación los métodos no jerárquicos para la formación de clústeres. En este caso, el número de casos que pueden entrar en el análisis es elevado y por lo general no produce resultados gráficos semejantes a los del análisis jerárquico.

3.3. Métodos no jerárquicos para la formación de conglomerados

"Sí tú me odias, quedaré yo convencido
de que me amaste -mujer- con insistencia;
pero ten presente, de acuerdo a la experiencia,
que tan sólo se odia lo querido;
pero ten presente, de acuerdo a la experiencia,
que tan sólo se odia lo querido".

Ódiame

Rafael Otero (1921-1997)

Los métodos jerárquicos son operativos cuando el número de casos es relativamente pequeño. Cuando consideramos miles de casos deben buscarse estrategias que permitan formar los grupos o conglomerados mediante diferentes algoritmos que definan lo que es un grupo, y los criterios de distancia o similitud para pertenecer a él. Como hemos advertido, los métodos no jerárquicos son los adecuados cuando el número de casos es elevado. Vamos a considerar dos tipos diferentes de estimación de conglomerados. Uno de ellos más tradicional, el método de k-medias, en el que se debe indicar el número de clústeres a estimar y es aplicable exclusivamente a niveles de medición de razón o de intervalo. El segundo de los métodos se basa en el empleo de las medianas para vertebrar la formación de los clústeres[12].

12. Existen otros métodos como el análisis de clúster en dos pasos ("*two steps cluster analysis*"). Este método es aplicable a bases de datos con un gran número de casos, y no es imprescindible indicar un número previamente al análisis. Es decir, los propone automáticamente. Existe una cierta controversia en la literatura especializada sobre las condiciones de apli-

3.3.1. *Conglomerados mediante k-medias y k-medianas*

El análisis de clúster mediante K-medias o k-medianas es una herramienta diseñada para clasificar los casos en un número de grupos. Las características de los casos que pertenecerán a cada grupo no son conocidas previamente y se determinan a partir de las variables que se elijan. Es muy útil cuando el número de casos es elevado. Un buen análisis de conglomerados debe ser eficiente (determinando el menor número de clústeres posible) y eficaz, en la medida que ayude a construir tipologías y segmentos útiles y significativos. A diferencia de los métodos jerárquicos, los grupos que se construyen son excluyentes entre sí desde el inicio. El procedimiento intenta maximizar las diferencias entre grupos, buscando la máxima homogeneidad interna dentro de ellos.

Consideremos en primer lugar el método K-medias ("*K-Means cluster analysis*"), para determinar los clústeres. En este método, el nivel de medición de los datos debe ser de razón o de intervalo. Una diferencia importante es que en este tipo de análisis el investigador debe especificar cuántos grupos considera que existen previamente al inicio del análisis. En ese sentido, manteniendo su carácter exploratorio obliga al investigador a adoptar las decisiones que considere más adecuadas para optimizar la solución del número de clústeres. Mediante el método de k-medias, en primer lugar se selecciona un caso de referencia para cada clúster, que estén lo más separados posibles del centro de todos los casos. Posteriormente cada caso es asignado al grupo de cuyo centro se encuentra más próximo. A continuación el proceso intenta asignar cada caso a un clúster de forma que se reduzca la suma de cuadrados intra-grupos. Este procedimiento continúa hasta que la suma de cuadrados intra-grupos no puede ser reducida más.

Es evidente que los valores iniciales de cada clúster, sobre los que se van agrupando los casos, son muy importantes. Dado que los procedimientos no investigan todas las particiones posibles de los datos, siempre queda la posibilidad de otras particiones (grupos) que minimicen aún más la suma de cuadrados intra-grupos. Al operar sobre la base de minimizar la suma de cuadrados intra-grupos, los procedimientos basados en k-medias equivalen a un

---

cación y los resultados de este enfoque. (Johann Bacher, Knut Wenzig, Melanie Vogler, "SPSS Twostep Cluster – a first evaluation"). El procedimiento permite el empleo de variables con diferentes niveles de medición simultáneamente. Las simulaciones sugieren que las variables categoriales tienen un fuerte efecto en la formación de clústeres, imponiéndose sobre las de tipo intervalo. Otra cuestión importante es la dificultad para identificar las situaciones en las que no existen clústeres en los datos. Esta última es importante, dado que el procedimiento ofrece de forma automática un número de clústeres.

análisis multivariante de la varianza, donde los grupos (es decir, los casos que los conforman) no son conocidos previamente. Por esta razón, aún cuando empleemos el análisis de varianza para evaluar la validez de los grupos, es importante recordar que el procedimiento se orienta a optimizar el valor de F por lo que es fácil que produzca valores elevados.

Es importante considerar que con este método partimos de un número concreto de clústeres. El procedimiento por el cual se van construyendo los clústeres varía dependiendo de si se conoce el valor del centro de los grupos, o si por el contrario los centros deben de ser estimados de forma iterativa, eso sí, siempre partiendo de un número prefijado de clústeres.

Ciertamente no siempre es posible tener una idea clara de cuántos grupos pueden definir las distancias. Por eso, dado que este procedimiento exige que se le indique el número de grupos que debe calcular, una solución es extraer una muestra aleatoria del total de los datos y sobre esta muestra parcial efectuar un análisis jerárquico. Tal y como hemos visto anteriormente. Tras efectuar esa aproximación de forma exploratoria, se puede determinar aproximadamente cuántos grupos parecen estar presentes en la matriz de datos y, asimismo, mediante este análisis exploratorio previo es posible estimar un valor inicial para el centro de cada clúster. Los valores que corresponden en el análisis jerárquico con los grupos buscados serían los valores de partida para construir los k-grupos. El valor inicial para formar los clústeres a partir de él es un elemento importante que debe decidir el investigador.

Ya con estos datos preliminares, el número de grupos y el valor de sus centros, es posible iniciar el ajuste de los casos a dichos grupos mediante el análisis de k-medias. Consideremos este caso en el que los centros de los grupos son conocidos. Para cada caso calcularemos su distancia al centro de cada uno de los grupos. El caso será agregado al grupo de cuyo centro se encuentre más próximo. Lo ideal es que la solución final de clústeres, muestre unos grupos cuyos centros se encuentran muy separados entre sí, y donde además los casos que pertenecen a cada grupo se encuentren muy próximos a su centro. Este es un dato importante para determinar la bondad de la solución.

Otros métodos alternativos para estimar el centro de los clústeres analizan los datos varias veces. Debemos partir de la idea de que una buena solución de clústeres debe de separar los casos lo bastante bien. Para ello, una estrategia posible es partir de los casos con una mayor distancia entre ellos y tomarlos como una estimación de los centros de los futuros clústeres. Así, se tomarían tantos casos como número de grupos se haya especificado inicialmente. Conforme se van incorporando más casos, uno de ellos ocupará

el centro del conglomerado cuando su distancia más corta a uno de los centros sea mayor que la distancia entre ese centro con todos los demás.

Otras alternativas son tomar los k primeros casos (siendo k el número de clústeres) como centros iniciales para los grupos, o por el contrario, tomar los últimos k casos. También se pueden decidir de forma aleatoria los centros, eligiéndolos al azar, o en otra estrategia, agrupar aleatoriamente los casos en los k grupos, y calcular la media, o la mediana, según se esté procediendo, de los casos que forman cada grupo. Estas medias (o medianas según el caso), formarían los valores iniciales para ir formando los grupos.

Existe la posibilidad de efectuar un análisis de componentes principales y ordenar los casos según su valor en el primer componente. Después se dividen los valores por k (n/k), y se toma el primer valor de la primera partición como centro del primer clúster, el primer valor de la segunda partición de n/k como centro del segundo clúster, y así hasta tomar todos los primeros valores de cada partición.

Veamos el siguiente ejemplo, considerando los datos de calidad democrática, podemos observar como los valores iniciales son los más extremos. En una solución de dos conglomerados, la DIM1 inicia el clúster 1 con el valor 8 (el máximo de esa variable) y el clúster 2 con el valor 1 (el mínimo de esa variable).

Centros iniciales de los conglomerados

	Conglomerado	
	1	2
DIM1	8,00	1,00
DIM2	9,50	3,80
DIM3	,70	-2,30
DIM4	1,30	-1,70

Como sabemos el procedimiento continúa de forma iterativa incorporando cada unos de los casos según su distancia. El historial de iteraciones nos muestra como se producen cambios muy importantes en las dos primeras fases, y después el cambio es muy lento hasta alcanzar la convergencia.

Historial de iteraciones[a]

Iteración	Cambio en los centros de los conglomerados	
	1	2
1	3,433	3,544
2	,572	,253
3	,095	,018
4	,016	,001
5	,003	9,225E-5
6	,000	6,589E-6
7	7,359E-5	4,707E-7
8	1,226E-5	3,362E-8
9	2,044E-6	2,401E-9
10	3,407E-7	1,715E-10
11	5,678E-8	1,225E-11
12	9,463E-9	8,758E-13
13	1,577E-9	6,172E-14
14	2,629E-10	5,032E-15
15	4,381E-11	9,222E-16
16	7,302E-12	1,110E-16
17	1,217E-12	,000
18	2,033E-13	,000
19	3,390E-14	,000
20	5,626E-15	,000
21	9,155E-16	,000
22	,000	,000

El resultado de esta iteración es una nueva estimación del valor de cada variable respecto al centro de cada conglomerado. Este centro final se calcula como la media para cada variable en el conglomerado final. En cierto modo, expresa los valores característicos de un caso típico en cada clúster.

Centros de los conglomerados finales

	Conglomerado	
	1	2
DIM1	6,91	4,19
DIM2	5,72	3,28
DIM3	1,90	-,73
DIM4	1,08	-,42

Al igual que sucederá en el análisis factorial, el perfil de las variables en cada uno de los conglomerados permite describir el segmento. Es decir, qué rasgos caracterizan a los que pertenecen a dicho grupo.

La tabla donde se muestra los centros de los grupos no facilita, sin embargo, información alguna respecto a la consistencia interna de los grupos. Por eso puede resultar conveniente efectuar un análisis de varianza, con los grupos como factor y cada una de las variables empleadas en el análisis (para estimar las distancias) como dependiente.

La media cuadrada entre clústeres se etiqueta en la columna "Conglomerado media cuadrática" y la media cuadrada intra-grupo se etiqueta "Error media cuadrática". El ratio entre ambas es el que aparece en la columna F. Un ratio F elevado y una significación baja indican que las variables son muy diferentes en sus valores para los diferentes clústeres. En todo caso, este test solamente es útil a efectos descriptivos. No lo es para testar la igualdad de las medias entre grupos, dado que el procedimiento empleado ha intentado optimizar ese efecto. Sin embargo es útil para conocer que variables tengan más peso e influencia en la solución. En el caso de la calidad democrática, se ha pedido que genere dos clústeres.

ANOVA

	Conglomerado		Error		F	Sig.
	Media cuadrática	gl	Media cuadrática	gl		
DIM1	24,272	1	1,960	16	12,383	,003
DIM2	34,614	1	1,912	16	18,104	,001
DIM3	19,433	1	1,821	16	10,674	,005
DIM4	3,240	1	1,248	16	2,597	,127

En el cuadro podemos apreciar como las diferencias de los valores de las variables DIM1, DIM2 y DIM3 son significativas en los dos clústeres. La significación de DIM1 es de .003, la de DIM2 es .001 y la de DIM3 es de .005, todas ellas por debajo de .05. Sin embargo en el caso de DIM4 la significación es de .127 expresando que la variable no es probablemente diferente en los dos clústeres. La bondad de la solución final se estima nuevamente según la capacidad que tengan los grupos para discriminar entre los valores de los casos en las variables consideradas.

Un dato importante es la distancia final entre clústeres. Cuanto más distanciados, más seguridad se tiene de que los segmentos o tipologías representan grupos con características diferentes.

Distancias entre los centros de los conglomerados finales

Conglomerado	1	2
1		4,744
– 2	4,744	

Por último, se ofrece información sobre cuántos casos existen en cada conglomerado (tipología o segmento).

Número de casos en cada conglomerado

Conglomerado	1	5,000
	2	13,000
Válidos		18,000
Perdidos		,000

El análisis de clúster es una técnica eminentemente exploratoria, y serán sucesivos diagnósticos los que ayudarán a perfilar y decidir los grupos más significativos.

También es posible aplicar el análisis k-grupos empleando medianas en lugar de medias. El procedimiento es esencialmente igual que para k-medias, excepto que se emplea la mediana para reasignar cada caso a cada clúster, y que el criterio de referencia es minimizar la suma intra-grupos de las desviaciones absolutas.

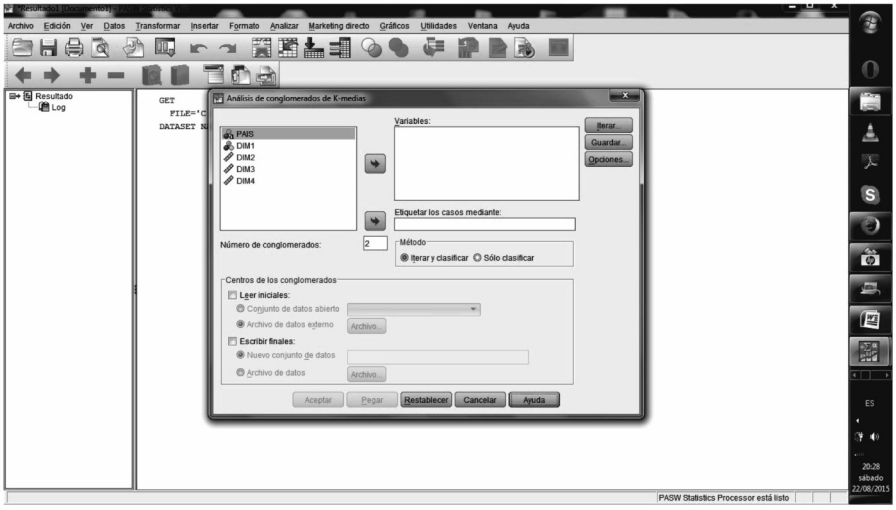
Tanto los procedimientos jerárquicos como los basados en k-grupos son los procedimientos más utilizados en la detección y estimación de clústeres

cuando consideramos la investigación en Ciencias Sociales. Esta área de actividad, detectando grupos y clústeres, dado su elevado interés en los nuevos procedimientos de "big data" y "minería de datos", experimenta en la actualidad un desarrollo bastante intenso.

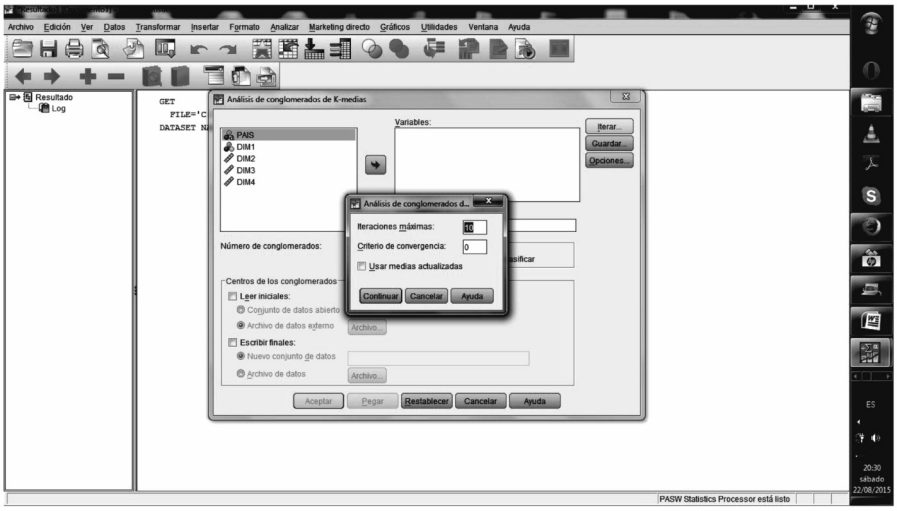
3.3.2. *Ejemplos en SPSS y SYSTAT*

En SPSS, tras seleccionar las variables o indicadores que se van a utilizar para definir los clústeres, se indica el número de grupos que consideramos que existe en los datos. Tal y como se comentó anteriormente, se puede proponer un centro de grupo para comenzar el procedimiento de conglomeración. Asimismo, se puede elegir entre dos formas de clasificar los casos entre los diferentes grupos: clasificando y recalculando los centros de los grupos conforme avanza el proceso, o simplemente clasificando los casos. Como resultado del análisis, se puede guardar como nuevas variables el clúster de pertenencia de cada caso, su distancia al centro y el valor final del centro del clúster.

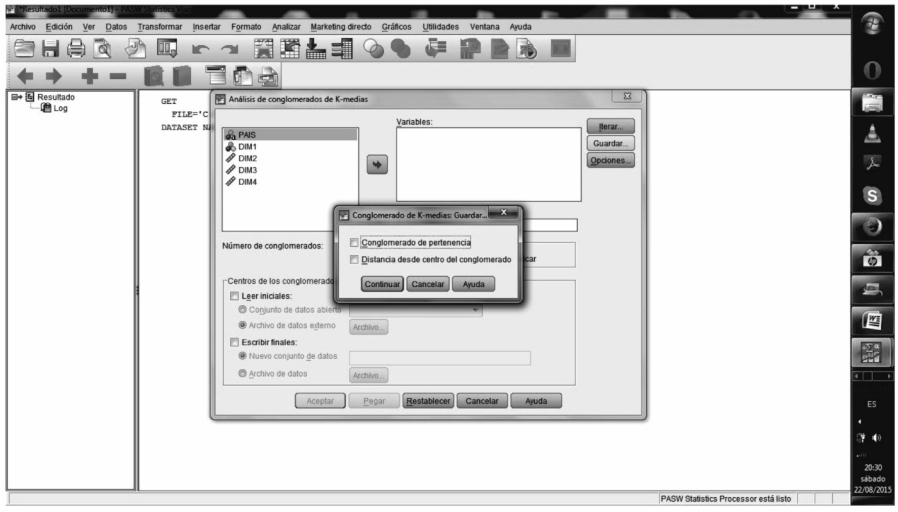
Como información para validar el significado de la solución, se puede solicitar que se efectué un análisis de la varianza. Ya se relativizó anteriormente el significado del valor F, dado que el procedimiento de estimación de clústeres está orientado a optimizar su valor, pero no obstante puede facilitar información importante en relación al peso o influencia que tiene cada variable en la separación entre grupos. Recordemos que el significado de los diferentes grupos (llamémosles clústeres, tipos, segmentos, etc.) depende de las variables que los definen.



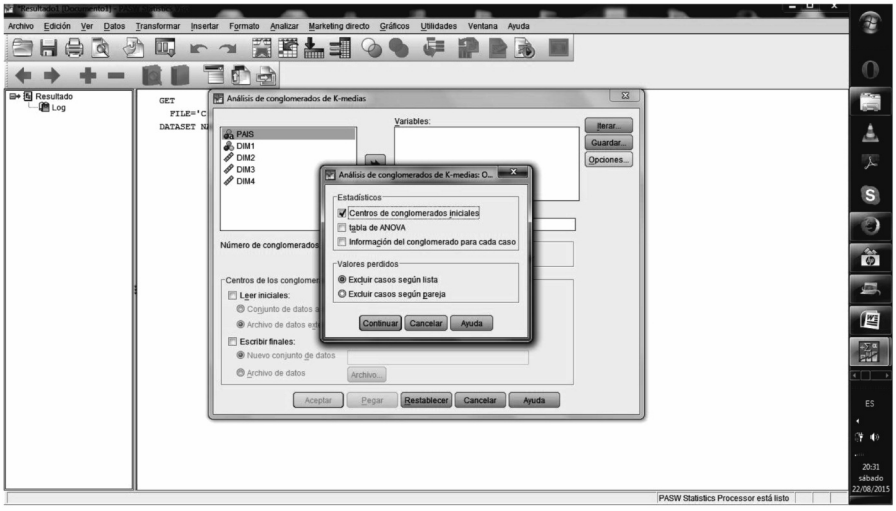
Este programa permite, asimismo, indicar el número máximo de iteraciones así como el criterio de convergencia. Lo habitual es mantener el criterio de convergencia en su valor 0, y elevar el número de iteraciones en el caso que no se alcance.



En la opción "guardar" se indica que se produzca una nueva variable con el grupo al que pertenece cada caso, así como la distancia a su centro de clúster.



En "opciones" se solicita la información relativa a los centros que se emplean para iniciar la conglomeración (construcción de los grupos), el análisis de varianza y la información del conglomerado para cada caso. Esta última información producirá, habitualmente, una información muy extensa, dependiendo del número de casos.



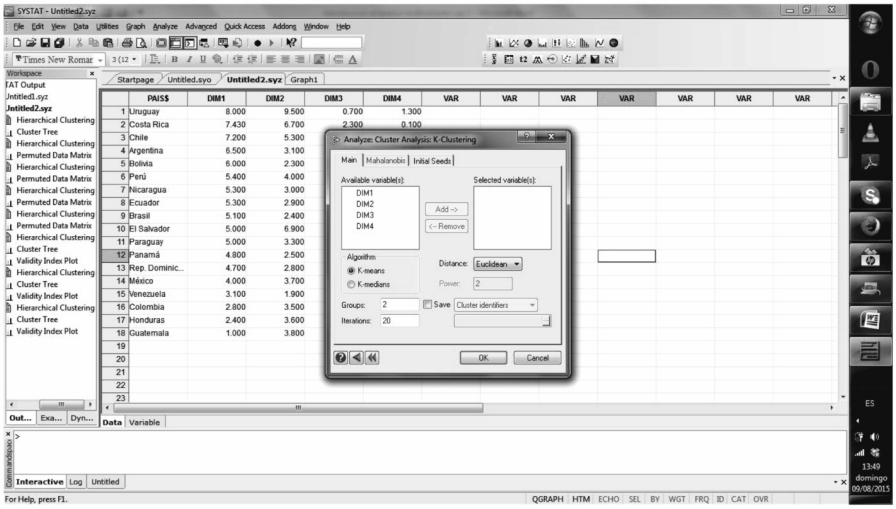
Finalmente, permite dos criterios para excluir casos del análisis según sus valores perdidos. Excluir los casos que tengan algún valor perdido en alguna variable (lista), o excluirlos parcialmente de aquellas parejas de variables en las que presente un valor perdido. En esta situación, los casos aparecen o desaparecen según su valor en cada pareja.

Clústeres con K-medias y K-medianas en SYSTAT

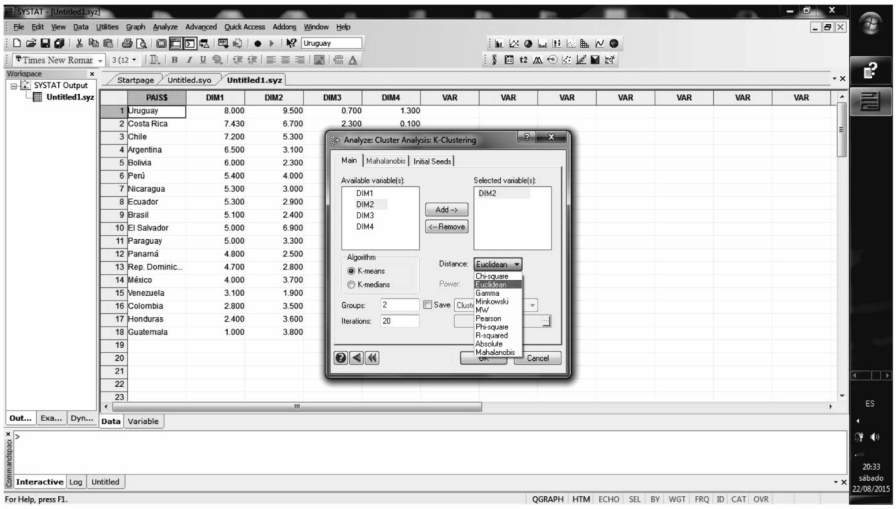
En SYSTAT se pueden realizar clústeres mediante K-medias y K-medianas. Los dos procedimientos tienen el mismo objetivo, maximizar las diferencias entre grupos y optimizar la homogeneidad intragrupos. En cierto sentido, equivale a efectuar un análisis de varianza donde se desconocen los grupos y se procede reclasificando de forma que el valor F se optimice.

En la ventana principal se eligen las variables, como es usual. Se debe elegir qué procedimiento se desea utilizar para la formación de clústeres, la media o la mediana, método más robusto a los casos extremos. Seguidamente se debe indicar el número de grupos que se quiere investigar. El número por defecto es dos. Se puede decidir el número máximo de iteraciones, con un valor por defecto de 20.

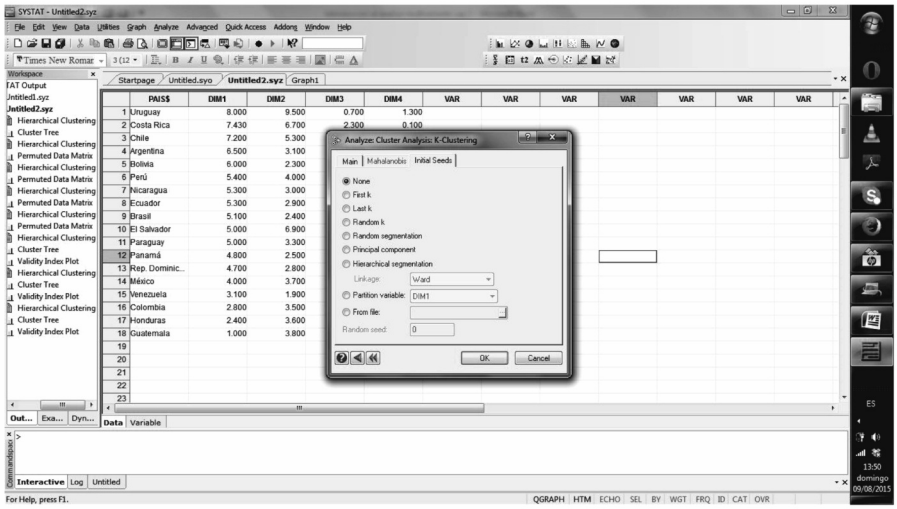
Debe decidirse qué distancia se va a utilizar para determinar las proximidades o las similitudes intra y entre clústeres.



Las distancias disponibles en SYSTAT para k-medias y k-medianas son Chi-cuadrado, Euclídea, Gamma, Minkowski, MW, Pearson, Phi-cuadrado, R-cuadrado, Absoluta y Mahalanobis. Es posible guardar en una nueva variable el grupo de pertenecía para cada caso así como los centros finales de cada grupo.



Por último, SYSTAT da nueve opciones para determinar cuáles van a ser los centros desde los que iniciar la agrupación de casos. "None" (ninguno) inicia el procedimiento con un grupo, y calcula su centro (media o mediana). A partir de él construye dos, basándose en el valor más alejado de ese centro, que pasa a ser el centro de un segundo grupo. Con esos dos centros procede a clasificar los casos de forma óptima. Continúa dividiendo grupos y reasignando casos hasta alcanzar el número de k-clústeres especificado. "First k" (primeros k casos), tras tomar los primeros k casos (que presenten valores válidos), los adopta como centros de inicio para clasificar el resto de los casos. "Last k" (últimos k-valores) emplea el mismo sistema, solamente que tomando los últimos k valores. "Random k" (aleatoria) elige de forma aleatoria los k centros para iniciar la clasificación. "Random segmentation" (segmentación aleatoria) construye k grupos de forma aleatoria y calcula sus respectivas medias o medianas. Posteriormente se adoptan dichas medias o medianas como valores iniciales para empezar a clasificar los casos. "Principal component" (componente principal) primero estima, y después toma, el primer componente principal como si fuese una variable. Tras ordenar todos los casos por su valor en el componente, divide el número de casos por k (número de clústeres) y construye los centros tomando el primer caso de cada grupo. "Hierarchical segmentation" (segmentación jerárquica), efectúa el procedimiento de clúster jerárquico, con el criterio de vinculación (linkage) que se especifique, y toma los k grupos que resultan de ese análisis como partición inicial que da origen a la clasificación. "Partition variable" (variable de partición) toma como cri-



terio para formar los k grupos (de los que extraer la media o mediana que inicia la clasificación) de una variable que especifique el investigador. "From file", la opción desde archivo indica que tenemos los centros de inicio de la clasificación escritos en un archivo externo. "Random seed" pide que los valores iniciales que deben referenciar la clasificación de los casos en los k grupos se estime de forma aleatoria.

4. EL ANÁLISIS FACTORIAL

Como se ha comentado anteriormente, los procedimientos estadísticos, en su mayor parte, son aplicables a diferentes objetivos. Otro ejemplo de esto es el denominado análisis factorial. En esencia, determina qué variables o indicadores están más próximos entre sí (forman clústeres) y partiendo de dichas agrupaciones, procede a estimar una puntuación para cada caso en ese grupo de variables. Al igual que en el análisis de clústeres, y como resultado de la agrupación de casos, se generaba una nueva variable donde se indicaba a qué clúster pertenecía cada caso, en el análisis factorial se emplean las variables agrupadas para calcular una puntuación para cada caso. La interpretación teórica es más amplia (los grupos de variables expresan un único concepto teórico, del que son expresión fragmentada), pero en la práctica, es una análisis de clústeres de variables que toma generalmente como unidad de proximidad la correlación o la covarianza.

Por ese motivo (empleo de la matriz de covarianzas o correlaciones), las variables deben ser de tipo cuantitativo, con niveles de medición de intervalo o razón. Como principio general, las variables para las que se pueda calcular el coeficiente de correlación de Pearson serían las más adecuadas. La razón básica es que el modelo de análisis factorial toma como presunciones que los datos deben de mostrar una distribución bivariable normal para cada par de variables y las observaciones deben ser independientes entre ellas. En la práctica existe un uso generalizado del análisis factorial en escalas de tipo Lickert, y en forma exploratoria, de variables dicotomizadas (aprovechando las posibilidades de la correlación tetracorica).

En sus aplicaciones concretas, es también un procedimiento multivariable para la construcción de índices. Este procedimiento permite integrar en un solo índice a un conjunto de indicadores o variables, siendo esta una de sus utilidades más relevantes. El análisis factorial representa una estrategia de medición amplia, útil para la exploración de conceptos teóricos, incluyendo el construir índices. En estas páginas desarrollaremos tanto su utilidad como herramienta para construir índices, como sus otras potencialidades analíticas.

Con frecuencia existen muchos conceptos en la investigación social que escapan a una observación directa. Pensemos por ejemplo en conceptos como alienación, anomía, poder, autoritarismo, xenofobia, racismo, etnocentrismo, etc. Resulta evidente que estas variables latentes no pueden medirse de una forma directa, tal y como puede medirse otras características como altura, peso, ingresos, género, etc.

Podemos considerar que estos conceptos se expresan a través de múltiples variables e indicadores. Así, sería el concepto "autoritarismo", el que podría explicar determinados comportamientos, opiniones, expresiones y actitudes de los individuos. Algo semejante puede decirse de conceptos como etnocentrismo, xenofobia o racismo. De hecho, el análisis factorial entronca en la labor de los psicólogos para determinar conceptos no medibles directamente, como puedan ser las formas diferentes de inteligencia o los estilos cognitivos. El concepto análisis factorial es propio de la terminología psicológica. En otras disciplinas se denominan también como "variables latentes" o "dimensiones".

En ese sentido, la respuesta que buscamos en los procedimientos estadísticos, es identificar qué grupos de variables están estrechamente relacionadas entre sí, postulando que esa estrecha asociación entre ellas responde a la existencia de un factor (dimensión o variable latente) que no es observable directamente. Para ello, es evidente que se debe descartar teóricamente la posible existencia de una secuencia explicativa entre las variables observadas. En ese caso, la estrecha asociación puede estar expresando una secuencia argumental (explicativa) y no un concepto. Por eso, las variables observadas (o los indicadores) deben teóricamente formar parte de una batería de preguntas o a un escalamiento de medición. El riesgo de tomar dependencia por medición puede aparecer, por ejemplo, en el análisis de datos secundarios donde se desconoce la intencionalidad original del investigador.

Evidentemente, es fundamental que los factores sean significativos. Tengan un significado teórico claro. Especialmente, esta situación se plantea cuando se efectúan análisis factoriales exploratorios, donde no se propone un concepto teórico que dé cuenta de la variabilidad observada. Nos encontraríamos en la ocasión de descubrir conceptos no conocidos, o no previstos. En el caso que la estructura sea significativa empírica (tras ser testada con nuevos datos) y teóricamente, se debe asignar un núcleo teórico que le aporte un significado sustantivo. El descubrimiento de un factor significativo, teórica y empíricamente, equivale a descubrir un nuevo mundo analítico, dónde pueden producirse nuevas ideas o planteamientos. No es estrictamente el caso (no está definido como resultado de un análisis factorial) pero un ejemplo puede ser el de los conceptos "materialismo" o "posmaterialismo" de Inglehart,

el de "incertidumbre" en Beck o el de "reflexividad" en Giddens, "Ideología política" (Alaminos, 2004), entre otros. En el caso que nos ocupa en Ecuador y el proyecto de medición del "Buen vivir", ya existe una reflexión teórica producto de un intenso debate académico e investigador, sobre los factores que podrían dar cuenta de esa realidad social.

El análisis factorial es especialmente útil en su función exploratoria, donde se agrupan las variables con mayor correlación entre ellas, para construir otras variables denominadas "factores" de tal modo que unas variables tengan una correlación mayor con unos factores, y prácticamente nula con otros. El carácter exploratorio significa que lo habitual es efectuar varios análisis, cambiando criterios como pueden ser el método de ajuste, el tipo de rotación, el criterio de selección de factores, hasta revelar la estructura que pueda existir en los datos. El significado teórico de los factores se obtiene mediante el significado de las variables que le dan forma, mostrando una elevada correlación con ellos.

En esa contigüidad que se aprecia entre muchos de los procedimientos estadísticos, debemos destacar como procedimiento "gemelo" al Análisis Factorial el Análisis de Componentes Principales. Aunque ambos procedimientos se basan en modelos matemáticos diferentes, se pueden utilizar con el mismo tipo de datos, produciendo además, por lo general, resultados muy parecidos.

La realización de un análisis factorial puede tener varios objetivos. Como hemos comentado reiteradamente, un procedimiento estadístico (respetando sus presunciones) puede ser aplicado con diferentes utilidades según los intereses del investigador. En esta ocasión, nos concentraremos en tres de sus utilidades más relevantes: la construcción de índices, la determinación de la estructura dimensional que pueda estar presente en un conjunto de variables o indicadores, y la medición de variables subjetivas.

a) La estimación de un índice, expresando de forma sintética un conjunto de indicadores. En esta situación se opera, usualmente, en un plano confirmatorio. La medición está establecida teóricamente y el procedimiento estadístico se utiliza como fórmula instrumental para construir el índice.
b) También, por ejemplo, otro objetivo es representar de forma armoniosa y simplificada un conjunto de variables. Para ello, se aspira a sintetizar un máximo de correlaciones observadas entre variables, empleando el mínimo número de factores posibles. En ese sentido, un exceso de factores implicaría muy poca simplificación. Cabe recordar aquí que la aspiración última es revelar una estructura latente con significado

teórico sustantivo. En ningún caso es el objetivo reducir el número de factores (simplificación) sacrificando interpretación teórica.

c) El análisis factorial de medición corresponde, si lo consideramos desde el punto de vista de la psicología, con un test psicométrico que mediría la presencia en los individuos de rasgos psicológicos (por ejemplo "personalidad autoritaria"), así como su intensidad. Siguiendo esa lógica, este tipo de análisis es aplicable a cualquier caso donde intentemos proponer la presencia de un concepto no medible directamente, como una realidad que se expresa mediante una serie de indicadores observados. En ese sentido el análisis factorial, como definición multivariable, es una técnica estadística que se emplea para identificar y medir un grupo pequeño de factores que dan cuenta de la relación que se aprecia entre un número más elevado de variables. Karl Pearson, quien desarrolló la técnica del análisis de componentes principales en torno al 1901, definía un "componente" como la línea que mejor se ajusta a sistemas de puntos en el espacio. Así, mientras que la idea tras la recta de regresión es la predicción, en el caso del análisis de componentes principales la idea clave es la de asociación.

Considerando la dimensionalidad de la solución factorial (cuántos factores y su relación con las variables o indicadores), es importante que se diferencie entre la identificación de un número de factores que simplifica el volumen de variables observadas, de la posibilidad de que dichos factores puedan o no estar correlacionados entre sí. Éstos son dos conceptos diferentes, el de simplificación de las variables observadas por un lado, y el de la relación existente entre los factores por otro. En el caso del análisis de componentes principales los componentes (factores) siempre van a ser independientes entre ellos (ortogonales). Por el contrario, en el análisis factorial se puede especificar que los factores sean interdependientes entre ellos (oblicuos).

Debemos diferenciar entre factores (también denominados dimensiones o variables latentes) de primer orden y factores (variables latentes o dimensiones) de segundo orden. La lógica es la misma. Los factores que se identifican y miden mediante la asociación entre variables observadas (o indicadores), pueden, a su vez, estar asociados entre ellos, expresando la presencia de un factor (variable latente o dimensión) de segundo orden. Es decir, la asociación entre variables observadas permite establecer la existencia de un factor, no visible directamente, que causa su variabilidad. En ese sentido, también es factible que los factores de primer orden (gracias a su posible asociación entre ellos) expresen la existencia de un factor de segundo orden, más profundo y solamente observable mediante el comportamiento de los factores de primer

orden. En ese sentido, la parsimonia se aplica en primer lugar reduciendo el número de variables gracias a los factores, y en algunos casos, reduciendo el número de factores mediante su agrupación en nuevos factores. Este proceso ayuda a simplificar la complejidad de la realidad que observamos y medimos directamente.

Un ejemplo de esto anterior es el concepto de "Buen Vivir". En el proceso para medirlo, y tras debatir lo que pueda significar el buen vivir, se identifican factores (dimensiones, variables latentes) que puedan agrupar su expresión en la sociedad. Con ello, el proceso para medir el buen vivir se simplifica notablemente si es posible identificar una serie de factores que expresan lo que significa. De hecho, el proceso de medición y definición podría, con facilidad, llevar a una dinámica de simplificación relacional que determine el buen vivir como un factor, no ya de segundo orden, sino de tercer o cuarto orden. A la fecha de escribirse este manual, y de forma exploratoria, el debate identificaba cinco amplias dimensiones (factores) de referencia: Democracia y participación, Movilidad, Inclusión social y derechos, Medio Ambiente y Economía popular. No obstante, estos factores pueden perfectamente ser de segundo o tercer orden, dependiendo de su forma operativa final.

4.1. El modelo matemático

El modelo matemático detrás de un análisis factorial, con varios factores, es bastante semejante a la ecuación de regresión múltiple. Por ejemplo, para el caso del indicador 1, este vendría expresado como el resultado de la combinación lineal de los diferentes factores propuestos. Ciertamente, la previsión es que las cargas sean más elevadas en unos factores que en otros. Consideremos el concepto "Buen vivir" como multidimensional, con cinco dimensiones: Democracia, Movilidad, Inclusión, Medioambiente y Economía.

$$\text{Indicador1} = a_1\,(\textit{Democracia}) + b_1\,(\textit{Movilidad}) + c_1\,(\textit{Inclusión}) + d_1\,(\textit{MAmbiente}) + e_1\,(\textit{EconomíaP}) + U_{ind1}$$

A diferencia del modelo de regresión múltiple, en este caso la dimensión *Democracia*, dimensión *Movilidad*, dimensión *Inclusión*, dimensión *MAmbiente y* dimensión *EconomíaP,* no son variables, sino que son los nombres que empleamos para referirnos al conjunto de variables que comparten ese concepto (y del que realmente son expresión). Esos grupos de variables son, como ya sabemos, los que definen el factor (o índice según objetivo del investigador). En ocasiones los factores que van a representar grupos de variables, no son conocidos, sino que deben de ser estimados empíricamente. En el caso

de la explicación del Indicador 1 anterior, las cinco dimensiones son denominadas factores comunes. De hecho, todos los indicadores (o variables) que se consideran en el análisis factorial pueden ser expresados como funciones de todos los factores, con un peso mayor o menor de cada uno de ellos.

$$\text{Indicador1} = a_1 (\textit{Democracia}) + b_1 (\textit{Movilidad}) + c_1 (\textit{Inclusión}) + d_1 (\textit{MAmbiente}) + e_1 (\textit{EconomíaP}) + U_{ind1}$$

$$\text{Indicador 2} = a_2 (\textit{Democracia}) + b_2 (\textit{Movilidad}) + c_2 (\textit{Inclusión}) + d_2 (\textit{MAmbiente}) + e_2 (\textit{EconomíaP}) + U_{ind2}$$

$$\text{Indicador 3} = a_3 (\textit{Democracia}) + b_3 (\textit{Movilidad}) + c_3 (\textit{Inclusión}) + d_3 (\textit{MAmbiente}) + e_3 (\textit{EconomíaP}) + U_{ind3}$$

Y así hasta el indicador n, cuando en el análisis se incluyen n indicadores (o variables observadas)

$$\text{Indicador n} = a_n (\textit{Democracia}) + b_n (\textit{Movilidad}) + c_n (\textit{Inclusión}) + d_n (\textit{MAmbiente}) + e_n (\textit{EconomíaP}) + U_{indn}$$

La letra U en la ecuación se denomina factor único, y representa aquella parte de la variabilidad que se observa en el Indicador (o variable) que no puede ser explicada por los factores comunes. Con carácter general la ecuación anterior puede expresarse de la siguiente forma.

Para una variable o indicador I

$$I_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 \ldots\ldots + A_{ik}F_k + U_i$$

Donde F son los factores comunes, la U es el factor único y las A son los coeficientes que combinan los k factores. Los factores únicos se asume que no están correlacionados entre sí y que tampoco están correlacionados con los factores comunes.

Cuando consideramos el análisis de componentes principales, observamos que la ecuación anterior se gira. Si el análisis factorial considera a los indicadores o variables observadas como el resultado de una combinación lineal de los factores más un error, el método de componentes principales considera los componentes como una combinación lineal de los indicadores o variables observadas.

Al igual que los indicadores (o variables) pueden expresarse como combinación lineal de los factores, los componentes son estimados empíricamente desde estos mismos indicadores (o variables observadas). En ese sentido la estimación de los componentes como combinación lineal de una serie de variables se notaría de la siguiente forma. Tomando, por ejemplo, la dimensión política,

esta sería el resultado de la combinación lineal entre las variables que se incluyen en el análisis.

$$\text{Democracia} = B_1 (\text{Indicador 1}) + B_2 (\text{Indicador 2}) + \ldots + B_n (\text{Indicador n})$$

donde las B son los coeficientes que relacionan los indicadores (variables) con el factor. En principio, es posible que todas las variables contribuyan al componente político en un mayor o menor grado, sin embargo, por lo general se espera que sea un conjunto de variables las que mayor impacto (B) tengan en dicho componente. La notación para estimar un componente J, (F_j),

$$F_j = W_{j1}I_1 + W_{j2}I_2 + \ldots.+ W_{jn}I_n$$

donde las W son llamadas puntuaciones factoriales, y n expresa el número de variables o indicadores.

Tal y como podemos apreciar, los factores son el resultado de combinaciones lineales entre los indicadores (variables observadas), y viceversa, es factible explicar la varianza de los indicadores (variables) mediante combinaciones lineales de los componentes o factores.

4.2. Diagnósticos de pertinencia del análisis factorial

Previamente a la realización de un análisis factorial, es importante evaluar primero la pertinencia de este tipo de análisis y, segundo, el grado de ajuste de la solución (número de factores y rotación) que se adopte. El gráfico siguiente muestra la secuencia seguida durante el análisis factorial. Desde la exploración de la matriz de correlación o covarianzas hasta el testado de las soluciones propuestas.

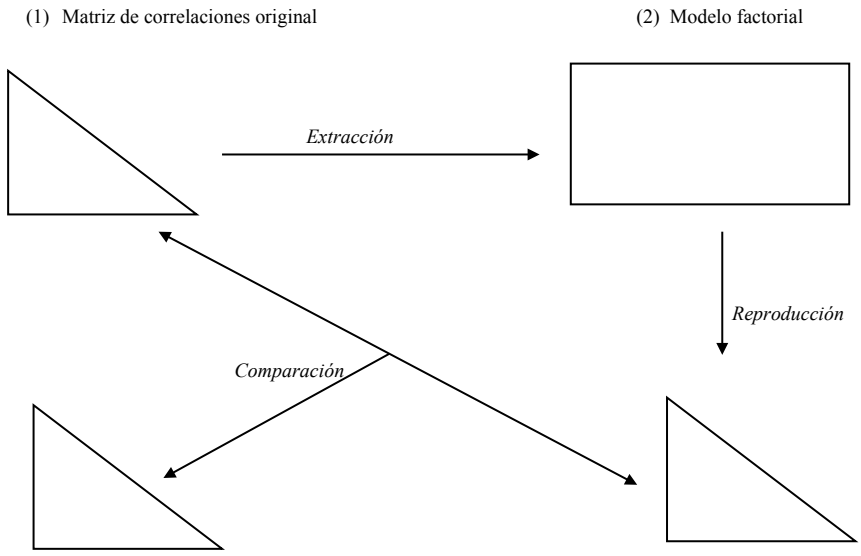
Las covarianzas se expresan normalmente en forma de matriz, por lo que el proceso consiste en descomponer[13] una matriz de covarianzas observadas en dos matrices: una matriz de covarianzas propuesta por el modelo de factores (reproducida), y una segunda que contiene los errores (diferencias entre covarianzas observadas y propuestas por el modelo). Sus diagonales son, precisamente, la comunalidad (covarianzas reproducidas) y la unicidad (errores en las covarianzas). En ese sentido, la ecuación básica del análisis factorial es

$$\text{Covarianzas observadas} = \text{Covarianzas entre factores} + \text{Error de las covarianzas}$$

13. Tanto el método de componentes principales, como el de máxima verosimilitud, al igual que otros empleados en el análisis factorial, son métodos para descomponer una matriz de correlación o de covarianza, desde la presunción de asociación (sin dependencia).

Es decir, que la covariación observada es consecuencia de la influencia de los factores, más un error aleatorio.

Cuadro Secuencia de un análisis factorial



En la fase (1) se evalúa la magnitud y relación entre los diferentes coeficientes de correlación. Así mismo, se comprueba que no es una matriz identidad. En la fase (2) se comprueba la comunalidad de cada variable, como consecuencia de la solución factorial elegida. En la fase (3) se reproduce la matriz de correlación desde el modelo factorial elegido. En la fase (4), tras comparar la matriz original de correlaciones y la reproducida, obtenemos los residuales. Unos residuales elevados en algún par de correlaciones indicarán que la solución factorial puede no ser adecuada para ellos. Así mismo, dará origen a estimar coeficientes de ajuste como KMO, CAM o la matriz y coeficientes AIC. Veamos este proceso más en detalle seguidamente.

Como sabemos, el concepto de factor implica que un conjunto de variables son la expresión de una dimensión latente. Por ello, los presupuestos del modelo postulan que las variables que expresan un factor deberían de estar altamente correlacionadas entre ellas. Si la correlación entre las variables es excesivamente baja, entonces difícilmente podríamos plantear que son la expresión de una realidad latente. Pero asimismo, la relación entre las variables que

forman un grupo (factor) y las demás debería mostrar una correlación baja. En definitiva, se espera existan clústeres de variables altamente relacionadas entre sí, y muy poco con las demás. Por ello, el análisis factorial se ocupa de descomponer la matriz de covarianza. La covarianza y la correlación son similares: la correlación es, en esencia, una covariación cuando las variables están normalizadas. Uno de los motivos para emplear la matriz de correlaciones y no la de covarianzas es para reducir el impacto de emplear variables con escalas muy diferentes. La correlación entre ingresos y edad es fácilmente comparable con la correlación entre otras dos variables con rango de 1 a 10, por ejemplo. Así, la matriz de correlaciones es útil cuando las variables están medidas en diferentes escalas, mientras que la matriz de covarianzas es preferible cuando el análisis se va a aplicar en múltiples grupos con diferentes varianzas en las variables consideradas.

Planteando un ejemplo, considerando lo anterior, una matriz de correlaciones como la siguiente, expresaría la posibilidad de la existencia de cuatro factores (o dimensiones) que serían los responsables de los grupos de variables correlacionadas entre sí. Esto es evidente en el caso de que las variables v1 a v12 representen una batería de variables o indicadores que intentan medir un fenómeno o realidad social.

Matriz de correlaciones simulada

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12
Var1	1											
Var2	**.9**	1										
Var3	**.7**	**-.8**	1									
Var4	**.9**	**-.7**	**.9**	1								
Var5	.1	.1	.2	**.9**	1							
Var6	.2	.2	.1	**.7**	**.6**	1						
Var7	.1	.1	.2	**.8**	**.7**	**.9**	1					
Var8	-.3	.2	.1	.2	.1	.2	**.6**	1				
Var9	.2	.1	.2	.1	.2	.1	**.8**	**.9**	1			
Var10	-.1	.2	.1	.3	.1	.2	**.7**	**.8**	**.6**	1		
Var11	.2	.1	.1	.1	.1	.1	.3	.1	.2	**.7**	1	
Var12	-.2	.1	.2	.2	.2	.1	.2	.1	.2	**.8**	**-.9**	1

En este caso que empleamos para ilustrar la idea anterior hay dos aspectos que deben advertirse. Primero, que las variables están ordenadas en la matriz según su correlación entre ellas, formando grupos. Esto no es habitual, y salvo

que el investigador decida construir la matriz con esa intención (introduciendo en el análisis las variables en un orden que forme grupos entre las variables que cree están relacionadas) lo más frecuente es que las variables estén desordenadas de modo que los grupos que forman son más difíciles de apreciar. Lo segundo, es que los coeficientes no acostumbran a ser tan diferenciados en su magnitud entre altos y bajos. Con frecuencia, alguna de las variables de un grupo está relacionada con un coeficiente de correlación significativo con alguna de las variables de otro grupo. Es decir, que parte de la variabilidad de una variable expresa un factor, y otra parte de su variabilidad expresa otro (u otros) factores.

En la investigación social es bastante frecuente encontrarnos con la situación donde una variable expresa varias dimensiones. Es en otras disciplinas, como la psicometría, donde el énfasis se pone en que una variable o indicador exprese un único factor o dimensión.

Por ello, una de las primeras acciones al plantear hacer un análisis factorial consiste en examinar la matriz de correlaciones entre las variables que van a formar parte del análisis.

Vamos seguidamente a considerar otro ejemplo donde los casos son los individuos que responden a una encuesta de opinión pública. Los datos proceden del Barómetro del Centro de Investigaciones Sociológicas, Estudio 3021 de abril de 2014. Las variables recogen las actividades de participación no convencional que han desarrollado los entrevistados.

V1 'Comprar ciertos productos por razones políticas, éticas o para favorecer el medio ambiente'
V2 'Dejar de comprar o boicotear ciertos productos por razones políticas, éticas o para favorecer el medio ambiente'
V3 'Participar en una huelga'
V4 'Asistir a una manifestación'
V5 'Asistir a una reunión o mitin político'
V6 'Contactar o intentar contactar con un/a político/a para expresarle sus opiniones'
V7 'Donar o recaudar fondos para una actividad social o política'
V8 'Contactar o comparecer ante los medios de comunicación para expresar sus opiniones'
V9 'Participar en un blog, foro o grupo de discusión política en Internet'
V10 'Firmar una petición/recogida de firmas'

La matriz de correlaciones siguiente expresa las relaciones entre las variables anteriores.

Matriz de correlaciones: variables de participación social

	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**V10**
V1	1									
V2	0,754	1								
V3	0,297	0,296	1							
V4	0,333	0,328	*0,715*	1						
V5	0,246	0,247	0,339	0,406	1					
V6	0,268	0,243	0,279	0,304	**0,47**	1				
V7	0,367	0,342	0,261	0,327	0,298	0,351	1			
V8	0,315	0,294	0,324	0,312	**0,345**	**0,521**	0,302	1		
V9	0,293	0,288	0,322	0,335	**0,281**	**0,406**	0,258	**0,48**	1	
V10	0,381	0,347	*0,405*	*0,462*	0,292	0,322	0,429	0,31	0,33	1

Se observan tres agrupaciones de variables con coeficientes de correlación elevados (entre ellas) y con coeficientes de correlación más reducidos con las demás variables. Como podemos apreciar, la magnitud de los coeficientes está lejos de ser tan elevados y tan homogéneos como los expuestos anteriormente a modo de ejemplo. Con carácter general, debería de apreciarse grupos de variables con una elevada correlación entre ellas, y simultáneamente estas variables deberían demostrar una correlación débil con las demás variables. Las agrupaciones de variables mostrarían una elevada correlación entre ellas (definiendo un factor), y muy baja con las variables que definen un factor diferente.

En términos estadísticos, podemos plantear una hipótesis nula según la cual la diagonal principal de la matriz de correlación estaría formada por unos, mientras que el resto de los términos fuera de la diagonal serían cero. En definitiva, se trata de comprobar que no nos encontramos ante una matriz identidad, en la medida que este tipo de matriz excluiría cualquier posibilidad de plantear al existencia de factores. La prueba estadística para comprobar que no se trata de una matriz identidad exige como requisito que los datos formen distribución conjunta multivariada normal. El test de esfericidad (basado en una transformación chi-cuadrado del determinante de la matriz de correlación) de Bartlett nos ofrece poder comprobar la hipótesis nula que afirma que la matriz de correlación pueda ser una matriz identidad. Cuando el coeficiente Chi-cuadrado del test de esfericidad es elevado, y la significación

asociada es baja podemos rechazar la hipótesis de que la matriz de correlación sea una matriz identidad. En el caso de que el coeficiente Chi-cuadrado de la prueba de esfericidad sea excesivamente bajo, cabe plantearse abandonar la idea de efectuar un análisis factorial de ese conjunto de variables.

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,819
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	8589,326
	gl	45
	Sig.	,000

Tal y como se observa para el caso que nos ocupa, el test de Bartlett ofrece un coeficiente chi-cuadrado de 8589.326 y una significación de 0.000. En ese sentido podemos rechazar la hipótesis nula que afirma que la matriz de correlación anterior pueda ser en realidad una matriz identidad.

Otro indicador de la fuerza de relación entre las variables es el coeficiente de correlación parcial. Cuando todas las variables comparten factores comunes, la correlación parcial entre pares de variables debería de ser baja cuanto se suprimen los efectos lineales de las demás variables. En definitiva, la correlación parcial es una estimación de la correlación entre los factores únicos. Estas correlaciones deberían de ser próximas a cero para poder cumplir las presunciones que habíamos planteado. Recordemos que se afirmaba que no existe correlación entre los factores únicos.

Una primera aproximación para emplear la información que aporta la correlación parcial es comparar la matriz de correlaciones parciales con la matriz de correlaciones observadas. Si la suma de los coeficientes de correlación parcial al cuadrado (entre todos los pares de variables) es muy baja, cuando se la compara con la suma de los coeficientes de correlación observada al cuadrado, el coeficiente será igual a 1. Es el denominado coeficiente Kaiser-Meyer-Olkin (KMO) de adecuación muestral. El coeficiente KMO expresa el sumatorio de correlaciones observadas al cuadrado, divididas por el sumatorio de las correlaciones observadas al cuadrado más el sumatorio de correlaciones parciales al cuadrado.

$$KMO = \frac{\sum \sum_{i \neq j} r^2_{ij}}{\sum \sum_{i \neq j} r^2_{ij} + \sum \sum_{i \neq j} a^2_{ij}}$$

Donde r_{ij} es el coeficiente de correlación simple entre las variables i y j, y a_{ij} es el coeficiente de correlación parcial entre las variables i y j. Valores bajos del coeficiente KMO indicarían que puede no ser una buena idea efectuar un análisis factorial, dado que la correlación entre pares de variables no puede ser bien explicada por las otras variables. Como orientación de los valores que alcanza el coeficiente KMO (Kaiser, 1974), se entiende que coeficientes entre 0.90 y 1.0 sería maravilloso, de 0.80 hasta 0.89 puede considerarse meritorio, mientras que 0.70 hasta 0.79 puede considerarse aceptable y 0.60 a 0.69 mediocre. Un coeficiente de 0.50 hasta 0.59 podría considerarse insuficiente, y ya por debajo de 0.49 sería inaceptable.

En el caso de la participación social que estamos considerando, el coeficiente de KMO es de .81, y podemos considerarlo como bastante bueno. En ese sentido, podemos continuar planteando la idea de que exista una lógica dimensional detrás de la disposición a la participación que muestran los entrevistados.

Al igual que se calcula un coeficiente para todos los pares de variables, es también posible calcular el coeficiente para cada variable individualmente. Este coeficiente se denomina Coeficiente de Adecuación Muestral (MSA por sus siglas en inglés).

$$CAM = \frac{\sum_{i \neq j} r^2_{ij}}{\sum_{i \neq j} r^2_{ij} + \sum_{i \neq j} a^2_{ij}}$$

Para ello, en la suma de cuadrados solamente se incluyen los coeficientes de correlación que incorporan la variable en consideración, como parte del par de variables. Los *coeficientes de adecuación muestral* se presentan en la diagonal de la matriz AIC. Nuevamente valores razonablemente elevados son convenientes para poder efectuar un análisis factorial. *Precisamente, en el caso de variables con coeficiente excesivamente bajo, podría considerarse la idea de suprimirlas del análisis factorial*.

Así mismo, otra estrategia posible para usar la información que aporta el examen de la correlación parcial, es el negativo del coeficiente de correlación parcial, que se denomina correlación anti-imagen, o AIC. La matriz anti-imagen es una buena indicación para testar si es apropiado o no un análisis factorial. Si en la matriz (fuera de la diagonal) existe un número excesivo de coeficientes elevados, habría nuevamente que abandonar la idea de efectuar un análisis factorial.

Matriz correlación anti-imagen

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1	,736[a]									
V2	-0,682	,730[a]								
V3	-0,006	-0,028	,775[a]							
V4	-0,026	-0,031	-0,603	,777[a]						
V5	0,006	-0,036	-0,029	-0,176	,883[a]					
V6	-0,02	0,026	0,003	0,012	-0,301	,834[a]				
V7	-0,094	-0,06	0,035	-0,068	-0,068	-0,132	,909[a]			
V8	-0,057	-0,026	-0,086	0,018	-0,056	-0,318	-0,046	,860[a]		
V9	-0,023	-0,051	-0,052	-0,06	-0,004	-0,147	-4,83E-05	-0,274	,900[a]	
V10	-0,102	-0,022	-0,08	-0,171	-0,008	-0,059	-0,237	-0,021	-0,087	,914[a]

a. Medida de adecuación muestral

Podemos observar en la diagonal valores elevados de KMO para cada variable, oscilando entre .73 (aceptable) y .90 (meritorio). Por la magnitud de los coeficientes de cada variable, no parece necesario retirar ninguna variable del análisis. Asimismo, el estudio de la matriz anti-imagen nos permite comprobar que los coeficientes son en general muy bajos, con alguno muy elevado, como entre V1 y V2, con un coeficiente de -0.6. De hecho la KMO de V1 y V2 son las más bajas de la diagonal. Es el momento de determinar la lectura teórica de las dos variables (comprar o no comprar productos), que expresan un cruce entre sociedad de consumo y posicionamiento ideológico y social. De hecho, la tercera variable en correlación con ellas se refiere al hecho de donar (dinero) a formaciones de orientación política. El analista debe incorporar la lectura al resto del análisis, sabiendo que el posible factor este contaminado por aspectos diferentes a la participación social, como es la orientación hacia el consumo.

Otra referencia a considerar es el coeficiente de correlación múltiple al cuadrado entre una variable y todas las demás. Es un buen indicador de la fuerza de la asociación lineal entre variables. Esos valores se mostrarán bajo la columna etiquetada "Comunalidades". Si el método de extracción es el de Componentes Principales, al inicio del análisis a cada variable se le concede una comunalidad de 1. Es tras la extracción que se determinará la comunalidad de cada variable en el contexto de la estructura factorial estimada. Aquellas variables con coeficientes de comunalidad bajos, (R^2 múltiple) serían buenas candidatas para ser eliminadas, en el proceso de optimizar la relación entre factores y variables.

En el caso que estamos considerando, es la variable que recoge la experiencia, o la intención, de "Donar dinero a una organización política" la que muestra una comunalidad menor (.38). Es decir, una menor relación lineal con las demás variables. No obstante, la cuantía de esta comunalidad no es decisoria por sí misma para excluir la variable del análisis.

Comunalidades

	Inicial	Extracción
Comprar ciertos productos por razones políticas, éticas o para favorecer el medio ambiente	1,000	,840
Dejar de comprar o boicotear ciertos productos por razones políticas, éticas o para favorecer el medio ambiente	1,000	,826
Participar en una huelga	1,000	,797
Asistir a una manifestación	1,000	,830
Asistir a una reunión o mitin político	1,000	,468
Contactar o intentar contactar con un/a político/a para expresarle sus opiniones	1,000	,703
Donar o recaudar fondos para una actividad social o política	1,000	,389
Contactar o comparecer ante los medios de comunicación para expresar sus opiniones	1,000	,639
Participar en un blog, foro o grupo de discusión política en Internet	1,000	,485
Firmar una petición/recogida de firmas	1,000	,485

Método de extracción: Análisis de Componentes principales.

Todos estos procedimientos que hemos considerado están orientados a determinar la consistencia entre la estructura de los datos y la estructura factorial (estructura latente) que proponemos para ella. Como hemos podido comprobar, aportan información sustantiva que va más allá de la estadística. Los coeficientes deben orientarnos sobre las decisiones que deben tomarse acerca de la inclusión o exclusión de variables en el análisis, e incluso sugerir ideas sobre la posible contaminación de otros significados en la varianza recogida por las variables, diferentes a los que pretendía la intención original de medición.

4.3. La estructura factorial

Una segunda labor importante es, partiendo de estos grupos de variables que se encuentran altamente correlacionadas entre sí, determinar la estructura de factores que mejor se ajusta sobre ellas. En este caso, establecer de forma empírica qué estructura factorial, y qué relación entre factores y variables, depende de lo que se conoce como métodos de extracción de los factores.

Existen varios métodos estadísticos y matemáticos para extraer soluciones factoriales de una matriz de correlaciones. Estos métodos se diferencian sobre todo por el criterio que emplean para definir un buen ajuste, entre los factores comunes y las variables correlacionadas. El método que se emplea por defecto en varios programas estadísticos, como el SPSS, (es decir, el método que se utiliza si no se indica nada al respecto), es el de Componentes Principales. En el análisis de componentes principales, se forman combinaciones lineales entre las variables observadas. El primer componente principal es la combinación que da cuenta de la mayor cantidad de varianza total. El segundo componente principal da cuenta de la segunda mayor cantidad de varianza, y no está correlacionado con el primer componente estimado y así sucesivamente se van estimado componentes que van explicando cada vez partes más pequeñas de la varianza total, y que además no están correlacionados entre ellos.

Es posible computar tantos componentes principales como variables haya. Si se utilizan todos los componentes principales, cada una de las variables está exactamente representada por ellos. Sin embargo, no habríamos producido ninguna simplificación dado que tendríamos tantos factores como variables observadas. Asimismo, cuando todos los componentes principales están incluidos, en la medida que ellos dan cuenta de toda la varianza, ya no habría necesidad de lo que denominamos como factor único, que en definitiva expresa la varianza no explicada de cada variable. Se explica esto anterior para una mayor comprensión del funcionamiento de la lógica de componentes principales. En este caso, lo habitual es conservar en el análisis como solución aquellos factores que cumplan una serie de requisitos (por ejemplo, un porcentaje significativo de varianza total explicada). Como consecuencia quedará una parte de la varianza total (y de cada una de las variables) sin explicar, dado que se efectúa una selección de factores.

En el caso del método de extracción basado en Componentes Principales, al iniciar el proceso la proporción de varianza explicada por los factores comunes, (comunalidad de la variable), es 1 para todas las variables. En general, el análisis de Componentes Principales es una técnica diferenciada del análisis factorial. Es decir, puede ser utilizada cuando se desea obtener combinaciones

lineales no correlacionadas, a partir de las variables observadas. Lo que hace en definitiva es transformar un conjunto de variables correlacionadas en un conjunto más pequeño de nuevas variables no correlacionadas (componentes principales). En el análisis factorial, en la actualidad, el método de extracción más frecuente es el de componentes principales. En este caso, integrado como método de extracción en el análisis factorial, tiene un empleo especial en la determinación de índices mediante métodos multivariantes.

No obstante, existen otras estrategias alternativas para estimar los factores. Como ya advertíamos, su diferencia fundamental es lo que se considera un "buen ajuste". Brevemente, el método de Factorización de Ejes Principales, procede de forma muy parecida al análisis de Componentes Principales, excepto que en la matriz de correlación la diagonal es sustituida por estimados de las comunalidades. En un primer paso, se emplean los coeficientes de correlación múltiple al cuadrado, como estimación inicial de las comunalidades. Basándose en ello se extraen un número de factores. Se vuelven a estimar las comunidades a partir de las cargas factoriales, y los factores son extraídos con la nueva estimación de comunalidad sustituyendo a la anterior. Éste proceso continúa hasta que dejan de producirse cambios significativos en la comunalidades. El Método de Mínimos Cuadrados No Ponderados produce, para un número fijo de factores, una matriz factorial que minimiza la suma de las diferencias al cuadrado entre las matrices de correlación observadas y la matriz de correlación reproducida (ignorando la diagonal). El Método de Mínimos Cuadrados Generalizados minimiza la diferencia anterior, sin embargo, las correlaciones son ponderadas de forma inversa a la unicidad de cada variable. Esto es, las correlaciones que implican las variables con una elevada unicidad reciben un peso menor que las correlaciones que implican variables con baja unicidad. El método de Máxima Verosimilitud produce una estimación de aquellos parámetros que muestran una mayor probabilidad de haber producido la matriz de correlación observada, si la muestra procediese de una distribución multivariada normal. Nuevamente las correlaciones son ponderadas por la inversa de la unicidad de las variables, mediante un algoritmo iterativo. El método Alfa considera que las variables que estamos empleando en el análisis son realmente una muestra de las variables potenciales. Mediante este procedimiento se intenta maximizar la fiabilidad de los factores. Con este método los autovalores ya no se pueden obtener como las suma de las cargas factoriales al cuadrado, y las comunalidades de cada variable no se determinan mediante la suma del cuadrado de las cargas factoriales en cada factor. Una exposición más extensa sobre los diferentes métodos de extracción puede encontrarse en Harman (1976); Mardia, Kent, y Bibby (1979) o Rencher (1998, 2002).

Son muchos los métodos alternativos disponibles para explorar la estructura (clústeres) de variables y proponer una reducción significativa (tanto teóricamente como empíricamente) de las variables o indicadores. No obstante, además de las referencias estadísticas, es esencial que el criterio fundamental que dirija la interpretación de la solución factorial sea de carácter teórico.

No obstante, además de la vertiente estadística y las diferentes concepciones de bondad de ajuste, un aspecto esencial es determinar cuántos factores necesitamos para representar los datos. Para ayudar a tomar esa decisión podemos considerar varios aspectos de tipo estadístico y teórico. Recordando siempre que el significado teórico es el más sustantivo, el estadístico es simplemente instrumental.

Desde el punto de vista estadístico, en el momento de decidir cuáles son los factores, es frecuente examinar el porcentaje de la varianza total explicada por cada factor. La varianza total es la suma de la varianza de todas las variables. Si en un análisis tenemos 20 variables, la varianza total sería igual a 20, dado que la varianza de cada una de las variables es 1. Para que sea más comprensible la lectura, tanto las variables como los factores se expresan de forma estandarizada, con una media de cero y desviación típica de uno. El total de la varianza explicada por cada factor aparece en la columna como autovalores. En ella se observa la varianza explicada que se le puede atribuir a cada factor. La última columna, muestra el porcentaje acumulado de la varianza explicada por cada factor, sumada con la varianza explicada de los que le preceden en la tabla. Normalmente los factores aparecen ordenados según la cantidad de varianza que explican.

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	4,172	41,724	41,724	4,172	41,724	41,724
2	1,212	12,119	53,842	1,212	12,119	53,842
3	1,076	10,761	64,604	1,076	10,761	64,604
4	,804	8,035	72,639			
5	,734	7,341	79,979			
6	,536	5,361	85,340			
7	,517	5,173	90,513			

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
8	,430	4,298	94,811			
9	,275	2,752	97,563			
10	,244	2,437	100,000			

Método de extracción: Análisis de Componentes principales.

En primer lugar aparece la varianza explicada (autovalores iniciales) por cada factor tras la extracción. Como podemos ver en el caso que nos ocupa de la participación, el primer factor da cuenta del 41,7% del total de la varianza, el segundo factor del 12,1% y el tercero del 10,7%. En conjunto, los tres factores explican el 64,6% de la varianza total. Son datos que se repiten en las columnas bajo la cabecera "*Sumas de las saturaciones al cuadrado de la extracción*".

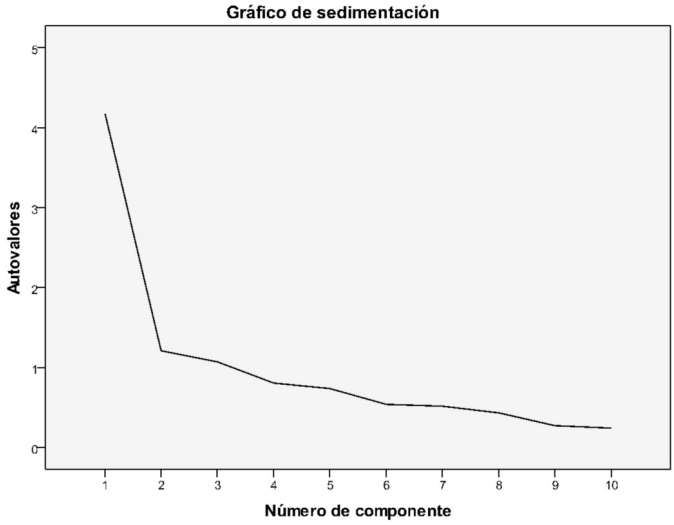
Hemos podido observar como el número de factores que permanecen en el análisis es objeto de una cierta controversia. En definitiva, desde el punto de vista estadístico, la idea es dejar fuera de la estructura factorial el mínimo posible de variación aleatoria. No obstante se han propuesto algunas reglas orientativas sobre cuándo parar de extraer factores como las siguientes (Dunteman, 1989: 22-3) (Box 6).

El criterio basado en Kaiser. Consiste en suprimir todos los factores con un valor eigen inferior a 1. Como sabemos, la principal razón es que no tiene sentido estadístico incluir un factor que explica menos que un indicador aislado. Partiendo de la varianza explicada, se han propuesto diferentes procedimientos para determinar el número de factores que deben ser empleados en un modelo. Algunos criterios sugieren que solamente aquellos factores que den cuenta de una varianza total mayor de 1 (autovalor > 1) deberían de ser tenidos en cuenta. La razón estadística es que aquellos factores con una varianza menor de 1 no son mejores que una simple variable, dado que cada variable tiene una varianza de 1. Pero esta es la lógica de la simplificación. Es evidente que esa varianza explicada por un factor, y que es menor de 1, tiene un origen y una composición diferente (la varianza total explicada) que la propia de una variable. Como sistema por defecto, el programa SPSS conserva aquellos factores con una varianza explicada superior a uno. Sin embargo, la solución de la simplificación no siempre es una buena opción, al olvidar el significado sustantivo que pueda reflejar un factor que explica poca varian-

za. En algunos casos, puede expresar la huella débil (consecuencia de no ser considerados originariamente en el diseño) de factores (variables latentes) no previstos, pero teóricamente significativos.

El uso del diagrama de sedimentación. Este método fue propuesto por Cattell y la recomendación es mantener los factores cuyos coeficientes caen abruptamente y eliminar aquellos que permanecen en un nivel semejante. Para ello se grafica de forma secuenciada la varianza explicada por cada factor (autovalores). En este tipo de gráfico se denomina de "sedimentación" y por lo general se aprecia algo parecido a una montaña, donde a partir de cierto momento la diferencia (en la varianza total explicada) entre los factores sucesivos es cada vez menor, dando la imagen de una ladera. Se entiende que la solución correcta, es decir el número de factores a conservar, son aquellos que muestran saltos importantes entre la varianza que explican y la explicada por el siguiente factor. Cuando la varianza explicada por los factores siguientes es poco significativa con respecto a los anteriores, se entiende que estos factores no son relevantes estadísticamente.

Considerando el ejemplo de la participación, podemos observar como el primer factor (denominado como "componente" en el gráfico), explica mucha más varianza que los demás, apareciendo muy distanciado. El segundo, aún con varianza total explicada superior a 1 está bastante más abajo y algo próximo al tercer factor, que también explica por encima de 1. A partir del cuarto factor podemos considerar que comienza la "ladera" con varianzas totales explicadas muy próximas entre sí.



El criterio de varianza explicada. En algunos casos, los investigadores mantienen en el análisis los factores que acumulan una varianza total explicada entre el 80% y el 90%.

El criterio de Joliffe. Consiste en eliminar los factores con un valor eigen por debajo del 0.70. La aplicación de esta regla puede hacer que se mantengan en el análisis casi el doble de factores que aplicando el criterio de Kaiser. En la medida que la simplificación es relevante en la mayoría de los casos esta regla es usada menos frecuentemente.

El criterio de comprensibilidad. Desde el punto de vista teórico y analítico, deberían retenerse aquellos factores que tienen un sentido teórico evidente. Los criterios de significado teórico y analítico se impondrían sobre los estrictamente estadísticos.

Sin embargo, las orientaciones estadísticas son útiles, pero no deben ni pueden suprimir la responsabilidad que tiene el investigador de decidir. Como sabemos, los dos objetivos principales del análisis factorial son simplificar, y mostrar una estructura con sentido teórico cuando la haya. La elección del *autovalor* o varianza explicada superior a 1, es un criterio estadístico, que sigue la lógica estadística. Sin embargo, debe primar ante todo el sentido teórico del investigador. Para ello, un aspecto muy importante es valorar la relación entre las variables y los factores.

4.4. La carga factorial

La relación empírica entre las variables y los factores se determina a través de la carga factorial. Esta se expresa en forma de matriz, donde en filas se indican las variables y en columnas los factores. Dependiendo del método de extracción utilizado, varía el nombre de la matriz. Si es el método de ajuste es por componentes principales, se denomina "matriz de componentes". En los otros casos, cuando se emplean métodos diferentes, es llamada matriz factorial. Al ejecutar un análisis factorial, se obtiene una primera solución en la que no se optimiza la relación entre variables y factores. En esa solución inicial, se expresa la relación de cada variable con los factores que han sido elegidos (ya sea por la magnitud de su autovalor o por criterios teóricos) para permanecer en el modelo. La tabla de la página 126 recoge la relación entre variables y factores en el ejemplo que estamos desarrollando.

En esta ocasión puede ser útil que recordemos la ecuación inicial donde se indicaba que se podían expresar cada indicador (o variables) como combinación lineal de los factores. Es algo que se puede apreciar con claridad en la matriz factorial. En la matriz factorial (o matriz de componentes) cada fila contiene los coeficientes para expresar la variable estandarizada en función

Matriz de componentes[a]

	Componente		
	F1	F2	F3
Asistir a una manifestación	,711	,122	-,556
Firmar una petición/recogida de firmas	,667	-,075	-,186
Participar en una huelga	,665	,147	-,577
Comprar ciertos productos por razones políticas, éticas o para favorecer el medio ambiente	,661	-,619	,141
Contactar o comparecer ante los medios de comunicación para expresar sus opiniones	,648	,280	,375
Dejar de comprar o boicotear ciertos productos por razones políticas, éticas o para favorecer el medio ambiente	,642	-,631	,126
Contactar o intentar contactar con un/a político/a para expresarle sus opiniones	,640	,377	,388
Participar en un blog, foro o grupo de discusión política en Internet	,613	,220	,246
Donar o recaudar fondos para una actividad social o política	,603	-,122	,105
Asistir a una reunión o mitin político	,601	,324	,039

Método de extracción: Análisis de componentes principales.
a. 3 componentes extraídos

de los factores. Esos coeficientes son llamados cargas factoriales, dado que indican cuanto peso se le asigna a cada factor en su relación con las variables. Los factores con coeficientes más elevados en valor absoluto expresan una relación más intensa entre esa variable y el factor. Por ejemplo, la variable asistir a una manifestación tendría la siguiente expresión

Asistir a una manifestación = .71 (F1) + .12 (F2) + (-.55) (F3)

El signo de la carga factorial (el coeficiente de cada factor para cada variable) en cada factor o componente es arbitrario. No lo es en la relación entre signos, que debe conservarse, pero sí su carácter positivo o negativo. En el caso que un factor o componente contenga más signos negativos que positivos, es factible cambiar el signo negativo a positivo, cambiando los positivos existentes a negativos. Algunos programas, como SYSTAT, realizan ese cambio de signos de forma automática cuando en un factor o componente hay más signos negativos que positivos. En ese caso, cambia los negativos a positivos y

viceversa. Este hecho hace que las soluciones factoriales que ofrece este programa puedan no coincidir en los signos de las cargas factoriales con las que ofrecen otros programas.

Los factores pueden estar correlacionados entre ellos o ser independientes entre ellos. En este segundo caso, cuando los factores estimados no están correlacionados entre ellos, se afirma que son ortogonales. Si los factores son ortogonales, las cargas factoriales expresan también las correlaciones entre los factores y las variables. La matriz de correlaciones entre las variables y los factores se denomina matriz de puntuación factorial. Dependiendo del tipo de rotación que pidamos, obtendremos tras la rotación una o dos matrices.

En el caso de rotación oblicua (factores interdependientes) se obtienen dos matrices, que se denominan "matriz de estructura" y "matriz de configuración". Cuando los factores son ortogonales, la "matriz de estructura" y la "matriz de configuración" son equivalentes y sólo se produce una única matriz que se denomina "matriz factorial".

Para efectuar una interpretación de la matriz factorial, tanto cuando los factores son ortogonales como cuando no lo son, podemos plantear que las cargas factoriales son los coeficientes de regresión estandarizados en la ecuación de regresión múltiple, donde la variable original es la dependiente y los factores las variables independientes.

Si además los factores no están correlacionados, los valores de los coeficientes no dependen unos de otros. Representan la contribución única de cada factor y definen la correlación entre factor y variable.

Cuando la rotación es oblicua, las cargas factoriales y las correlaciones entre las variables y los factores ya no coinciden. Las cargas factoriales continúan siendo los coeficientes de correlación parcial, pero ya no coinciden con la correlación (entre variable y factor). Los coeficientes de correlación se muestran en una nueva matriz denominada "matriz de estructura".

Así, en la rotación ortogonal se produce una sola matriz donde coinciden regresión parcial y coeficientes de correlación. En la rotación oblicua se producen dos diferentes, la "matriz de configuración" donde se recogen las cargas factoriales y la "matriz de estructura", donde se recoge la correlación entre factores y variables.

Para determinar cómo ajusta el modelo anterior de tres factores, y conocer cómo describe las variables originales, es posible calcular el porcentaje de varianza de cada variable que es explicada por el modelo de tres factores. Dado que en este ejemplo los factores no están correlacionados, la proporción total de varianza explicada es simplemente la suma de la proporción de varianza explicada por cada factor. Recordemos que la proporción de varianza explicada por los factores comunes determina la comunalidad de la variable.

Para calcular el porcentaje de varianza de una variable que viene explicada por los factores, se eleva al cuadrado el coeficiente de correlación entre factor y la variable.

$$\text{Varianza explicada de Participar en una huelga} =$$
$$(.665)^2 + (.147)^2 + (-.577)^2 = .44 + .02 + .33 = .79$$

Las comunalidades de las variables, aparecen en las estadísticas finales, tras mantener en el análisis el número deseado de factores. Las comunalidades pueden oscilar entre cero y uno. Cero indicando que los factores comunes no explican varianza alguna, y uno indicando que toda la varianza de la variable es explicada por los factores comunes. La varianza que no es explicada por los factores comunes se atribuye a lo que se denomina factor único o también unicidad de la variable.

Otra estrategia para conocer en qué condiciones se está ajustando el modelo, es mediante la matriz de correlación reproducida. Como sabemos una de las presunciones básicas del análisis factorial es que la correlación observada entre variables se debe a que comparten factores comunes. Por ello, la correlación calculada entre factores y las variables puede ser empleada para estimar las correlaciones entre variables. Es decir, reproducir las correlaciones sobre las que se ha construido el modelo.

4.5. Diagnóstico: La matriz de correlaciones estimadas (reproducidas)

Como sabemos, una vez especificado y ajustado el modelo de factores, podemos reproducir las correlaciones o covarianzas entre variables. En otras palabras, si la solución factorial ha logrado representar adecuadamente a la estructura de datos original (matriz de correlación), la que se genere desde el modelo debería parecerse mucho a la original. Para calcular y estimar las correlaciones entre las variables a partir de los factores, tomemos por ejemplo las variables A y B. Se multiplica el coeficiente (su carga) de la variable A por el coeficiente (carga) de la variable B en el primer factor, a ese resultado se le suma el producto de la carga de la variable A por la carga de la variable B en el segundo factor, y se le suma el producto de la carga de la variable A por la carga de la variable B en el tercer factor. La suma de productos es igual a la correlación estimada entre las dos variables. Es decir, desde el modelo (que simplifica la matriz de correlaciones observadas) se recalculan y reproduce la matriz de correlaciones que le dio origen.

Es posible solicitar que los programas nos impriman la matriz reproducida de correlaciones. A partir de ese momento, comparando la matriz de correlaciones observadas con la matriz reproducida de correlaciones, podemos obtener un residual para cada correlación comparada. Los residuales (es decir la

diferencia entre la correlación original y la reproducida) se muestran en la matriz de residuales, mientras que las correlaciones estimadas se expresan en la matriz de correlaciones reproducidas. En la diagonal aparecen las comunalidades. Los datos siguientes corresponden con el Barómetro de abril de 2014 del CIS, en sus preguntas sobre participación.

Correlaciones reproducidas

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1	,840a									
V2	0,832	,826a								
V3	0,267	0,262	,797a							
V4	0,317	0,31	0,812	,830a						
V5	0,202	0,186	0,425	0,446	,468a					
V6	0,244	0,222	0,257	0,285	0,522	,703a				
V7	0,488	0,477	0,323	0,356	0,327	0,38	,389a			
V8	0,308	0,287	0,256	0,287	0,495	0,666	0,396	,639a		
V9	0,304	0,286	0,298	0,326	0,449	0,571	0,368	0,551	,485a	
V10	0,461	0,452	0,54	0,569	0,369	0,326	0,392	0,342	0,346	,485a

Residual[b]

V1									
V2	-0,078								
V3	0,029	0,034							
V4	0,016	0,018	-0,097						
V5	0,044	0,061	-0,087	-0,039					
V6	0,024	0,021	0,021	0,018	-0,053				
V7	-0,122	-0,135	-0,062	-0,029	-0,028	-0,03			
V8	0,006	0,007	0,067	0,025	-0,15	-0,144	-0,093		
V9	-0,01	0,003	0,024	0,009	-0,168	-0,165	-0,11	-0,075	
V10	-0,08	-0,105	-0,135	-0,107	-0,077	-0,004	0,038	-0,031	-0,019

Método de extracción: Análisis de Componentes principales.

a. Comunalidades reproducidas

b. Los residuos se calculan entre las correlaciones observadas y reproducidas. Hay 21 (46,0%) residuales no redundantes con valores absolutos mayores que 0,05.

Debajo de la matriz aparece un mensaje indicando cuantos residuales son mayores de 0.05 en valor absoluto. Así por ejemplo, hay un 46% de residuales que

son mayores de 0.05. En ese sentido, tanto la cantidad de residuales mayores de 0.05 como la magnitud de los residuales indican en qué grado el modelo ajustado reproduce las correlaciones observadas. Si los residuales son grandes, el modelo no ajusta suficientemente bien y posiblemente deba ser reconsiderado.

Especialmente, puede evaluarse la relación por pares entre las variables para identificar aquellas que muestran una relación más débil. En definitiva, todos estos procedimientos ayudan a conocer más en profundidad el comportamiento de las variables e indicadores que se están considerando en el análisis.

4.6. Las rotaciones

Como hemos comentado anteriormente, ejecutar un análisis factorial implica en primer lugar, calcular la matriz de correlaciones o covarianzas a partir de la matriz de datos (casos x variables): Posteriormente, sobre dicha matriz se efectúa una primera estimación de las cargas factoriales. Es una primera extracción que no tiene en consideración la relación "teórica" entre variables y factores. Es tras esta primera extracción que el investigador puede solicitar que se optimicen determinados parámetros (imponiendo diferentes restricciones) de forma que facilite una mejor interpretación del significado de los factores o componentes. Por ejemplo, exigiendo que se optimicen las cargas de cada variable en cada factor, de forma que estas sean máximas o mínimas, pero no intermedias. Son varios los métodos que imponen restricciones en la solución factorial inicial que se extrae. Este proceso de imponer nuevas restricciones que refinen el resultado para una mejor comprensión teórica se denominan rotaciones. Entre los criterios más citados para la rotación destaca los propuestos por Thurstone. Así, para k variables y F factores o componentes, a) cada factor debería tener al menos F variables (tantas como factores) con cargas próximas a cero, y b) muy pocos factores deberían tener cargas elevadas en las mismas variables.

Las rotaciones son especialmente útiles para interpretar el significado de los factores. Como sabemos, las rotaciones no afectan a la varianza explicada ni a la comunalidad final. Continuando con el ejemplo se puede apreciar que, tras la rotación, la relación entre variables y factores permite identificar con mayor claridad el significado de los factores. El factor 1 (o componente) está claramente relacionado con la "comunicación", el segundo con la actividad "económica y de consumo" y el tercero con una "protesta más presencial y física".

Matriz de componentes rotados[a]

	Componente		
	1	2	3
Contactar o intentar contactar con un/a político/a para expresarle sus opiniones	**,823**	,110	,115
Contactar o comparecer ante los medios de comunicación para expresar sus opiniones	**,768**	,192	,113
Participar en un blog, foro o grupo de discusión política en Internet	**,641**	,197	,186
Asistir a una reunión o mitin político	**,574**	,065	,366
Comprar ciertos productos por razones políticas, éticas o para favorecer el medio ambiente	,157	**,891**	,147
Dejar de comprar o boicotear ciertos productos por razones políticas, éticas o para favorecer el medio ambiente	,130	**,887**	,147
Donar o recaudar fondos para una actividad social o política	,370	**,444**	,234
Asistir a una manifestación	,201	,176	**,871**
Participar en una huelga	,174	,126	**,866**
Firmar una petición/recogida de firmas	,273	,385	**,512**

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.
a. La rotación ha convergido en 5 iteraciones.

Las rotaciones las efectúan la totalidad de programas de análisis estadístico, siendo los métodos más habituales los siguientes. La rotación "Varimax", es una rotación ortogonal que intenta minimizar el número de variables que tienen una carga elevada en cada factor. Su utilidad facilita la interpretación teórica del factor o componente. La rotación "Quartimax" enfoca la cuestión de un modo diferente. Intenta minimizar el número de factores que se requieren para explicar cada variable o indicador. Con ello simplifica la interpretación de las variables o los indicadores observados. El método de rotación "Equamax" es una combinación de los dos anteriores, y persigue obtener una solución donde se minimice el número de variables con cargas elevadas en los factores y se minimicen el número de factores necesarios para explicar la varianza de cada una de las variables. Existe la posibilidad de controlar ese balance al que aspira la rotación "Equamax" mediante el método "Orthomax". El método "orthomax" controla la simplificación de variables y factores mediante el coeficiente

"gamma". Variando su valor modifica la optimización desde "varimax" a la optimización mediante "quartimax". Todos estos métodos son ortogonales, manteniendo la independencia entre factores.

El método "Oblimin" produce rotaciones no ortogonales. Es decir, que los factores no son independientes entre ellos. El grado de oblicuidad (correlación) entre los factores viene controlado por un coeficiente. Este coeficiente es diferente según qué programas se utilicen. Así, en SYSTAT el coeficiente "gamma" expresa el grado de correlación entre factores. Con un valor de cero se permiten correlaciones moderadas, cuando el valor positivo es más elevado se permiten correlaciones más fuertes. En el programa SPSS el coeficiente se denomina "delta". Cuando delta vale 0 la solución es la más oblicua. Conforme el valor va siendo más negativo la relación entre factores es menos oblicua. Otra alternativa es "Promax", que facilita una rotación oblicua y cuya ventaja esencial sobre "oblimin" es la rapidez de cálculo en grandes bases de datos.

Las rotaciones son herramientas para el investigador, muy útiles para conocer y explorar la consistencia de las soluciones. Forman parte de las decisiones y siempre deben estar indicadas en la solución final que se adopte. Por ello, tras la elección de cuántos factores retener en el análisis, es habitual proceder a la rotación de los factores para optimizar su significado, gracias a simplificar la relación de las variables con los factores. La suma de los valores eigen (autovalores) no se ve alterada tras la rotación, sin embargo los cambios en los ejes pueden alterar el valor eigen (autovalor) de algunos factores y modificar sus cargas factoriales.

Así, considerando el ejemplo anterior que estudiaba la estructura de la participación, podemos observar cambios tras la rotación. En el cuadro, en primer lugar aparece la varianza explicada (autovalores iniciales) por cada factor tras la extracción. Podemos ver en el caso que nos ocupa de la participación, que el primer factor da cuenta del 41,7% del total de la varianza, el segundo factor del 12,1% y el tercero del 10,7%. En conjunto, los tres factores explican el 64,6% de la varianza total. Son datos que se repiten en las columnas bajo la cabecera "*Sumas de las saturaciones al cuadrado de la extracción*".

En la segunda parte de la tabla, "*Suma de las saturaciones al cuadrado de la rotación*", podemos observar que la varianza total explicada es la misma, un 64,6%. Esto es así dado que el número de factores que se retienen continúan siendo tres. Sin embargo, la varianza total explicada atribuida a cada factor es diferente. La varianza total explicada por cada factor se ha redistribuido entre ellos, como consecuencia de las modificaciones que experimentan su carga factorial. La relación entre variables y factores se ha modificado y con ello la varianza que explica cada factor. El primer factor da cuenta tras la rotación del 23,3% del total de la varianza, el segundo factor del 20,6% y el tercero del 20,6%.

Varianza total explicada

Componente	Autovalores iniciales			**Sumas de las saturaciones al cuadrado de la extracción**			**Suma de las saturaciones al cuadrado de la rotación**		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	4,172	41,724	41,724	4,172	41,724	41,724	2,331	23,314	23,314
2	1,212	12,119	53,842	1,212	12,119	53,842	2,065	20,651	43,965
3	1,076	10,761	64,604	1,076	10,761	64,604	2,064	20,639	64,604
4	,804	8,035	72,639						
5	,734	7,341	79,979						
6	,536	5,361	85,340						
7	,517	5,173	90,513						
8	,430	4,298	94,811						
9	,275	2,752	97,563						
10	,244	2,437	100,000						

Método de extracción: Análisis de Componentes principales

Precisamente, una de las críticas a la rotación es que diferentes rotaciones producen diferentes cargas factoriales de las variables en los factores y con ello pueden producir diferentes significados para los factores en los que cargan. Tenemos que considerar que las variables permiten identificar el significado teórico del factor (en el análisis exploratorio especialmente), por lo que el cambio de carga de las variables influye en la posible modificación del significado del factor. No obstante, esta volatilidad potencial puede ser controlada, al menos en su presencia e impacto, comparando los efectos de diferentes rotaciones y el modo en que puedan afectar el significado de los factores.

Veamos seguidamente un ejemplo del empleo de las rotaciones buscando el significado teórico de la estructura factorial. Las variables siguientes analizan la confianza en instituciones en España, según el Barómetro del Centro de Investigaciones Sociológicas, de abril de 2014. En una escala de 0 a 10, se valoran las siguientes instituciones y actores sociales, políticos y económicos.

'*La Monarquía*'
'*El Gobierno*'
'*El Parlamento*'
'*El Gobierno de su comunidad autónoma*'
'*El Parlamento de su comunidad autónoma*'
'*El Tribunal Constitucional*'

'El Defensor del Pueblo'
'Las Fuerzas Armadas'
'La policía'
'La Guardia Civil'
'Los partidos políticos'
'Las organizaciones empresariales'
'Los sindicatos'
'Los medios de comunicación'
'La Iglesia católica'
'El Consejo General del Poder Judicial'

El estudio de los coeficientes que hemos considerado indica que es factible buscar una estructura factorial en los datos. Así, el KMO es de .90 (excelente) y el test de esfericidad de Bartlett indica un Chi-cuadrado de 19553,401 para una significación de .000.

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,904
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	19553,401
	gl	120
	Sig.	,000

Parece que es una idea sensata buscar una estructura factorial tras la valoración de la opinión pública de las instituciones. El análisis, tras la extracción mediante componentes principales, y valor de selección del autovalor igual o superior a 1 da lo siguiente.

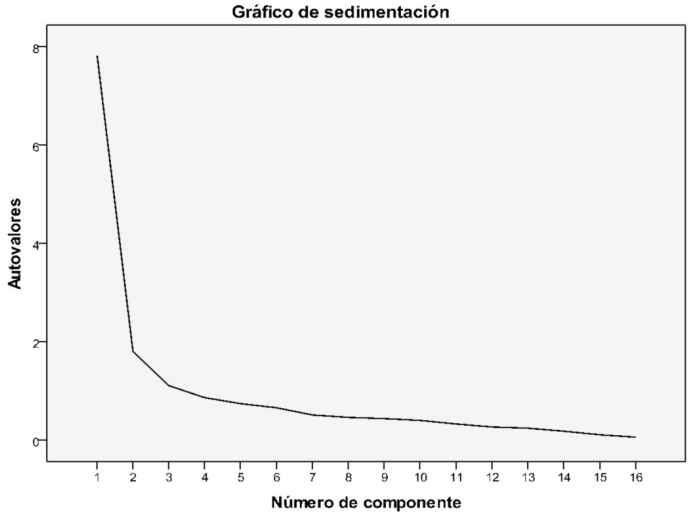
Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	7,825	48,906	48,906	7,825	48,906	48,906
2	1,810	11,312	60,217	1,810	11,312	60,217
3	1,110	6,935	67,152	1,110	6,935	67,152
4	,864	5,402	72,554			

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
5	,739	4,617	77,171			
6	,664	4,151	81,322			
7	,515	3,216	84,538			
8	,461	2,881	87,419			
9	,434	2,715	90,135			
10	,402	2,510	92,644			
11	,323	2,019	94,663			
12	,269	1,679	96,342			
13	,239	1,495	97,837			
14	,181	1,134	98,971			
15	,107	,670	99,641			
16	,057	,359	100,000			

Método de extracción: Análisis de Componentes principales

Considerando el gráfico de sedimentación, también sugiere que tres componentes (factores o dimensiones) pueden ser una solución adecuada.



Sin embargo, desde el punto de vista teórico, la identificación del significado de los factores es confusa al adoptar una estructura factorial de tres componentes.

Tras la rotación, en la matriz factorial siguiente podemos apreciar las cargas de cada variable en cada factor. Como sabemos, esto ayuda a facilitar la interpretación del significado teórico de los factores. Observando la siguiente matriz factorial es posible proponer algunas posibilidades.

Matriz de componentes rotados[a]

	Componente		
	1	2	3
El Parlamento de su comunidad autónoma	***,865***		,221
El Gobierno de su comunidad autónoma	***,856***		,201
El Parlamento	***,767***	,339	,228
El Gobierno	***,751***	,432	
Los partidos políticos	***,595***	,168	,486
El Tribunal Constitucional	***,520***	,442	,368
El Defensor del Pueblo	***,447***	,373	,413
La Guardia Civil		***,893***	,149
La policía	,131	***,870***	,192
Las Fuerzas Armadas	,174	***,855***	,132
La Monarquía	,491	***,587***	
La Iglesia católica	,428	***,527***	,174
El Consejo General del Poder Judicial	,415	***,477***	,455
Los sindicatos	,167		***,835***
Los medios de comunicación	,109	,255	***,654***
Las organizaciones empresariales	,429	,304	***,561***

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.
a. La rotación ha convergido en 4 iteraciones.

Los factores parecen ser "Democracia", "Autoridad" y un tercero de carácter "Social-económico". Sin embargo, explorando una solución de cuatro dimensiones y tras la rotación, se aprecia una estructura más definida. Al permitir un

cuarto factor, la dimensión "Autoridades" pasa a ser una de las dimensiones con más peso, según la varianza total explicada. La dimensión "Democracia" aparece desdoblada en dos tipos diferentes. Por una parte la democracia "representativa" con los partidos políticos o los parlamentos y por otra, el "poder judicial".

Matriz de componentes rotados[a]

	Componente			
	1	2	3	4
La Guardia Civil	***,896***		,162	,137
La policía	***,869***	,116	,187	,175
Las Fuerzas Armadas	***,825***	,119	,281	
La Monarquía	***,538***	,411	,365	
La Iglesia católica	***,537***	,427	,167	,168
El Gobierno de su comunidad autónoma		***,881***	,137	,217
El Parlamento de su comunidad autónoma		***,877***	,174	,226
El Gobierno	,387	***,674***	,390	
El Parlamento	,269	***,655***	,500	,117
Los partidos políticos	,135	***,533***	,373	,422
El Tribunal Constitucional	,281	,289	***,795***	,145
El Defensor del Pueblo	,225	,234	***,738***	,207
El Consejo General del Poder Judicial	,359	,241	***,651***	,286
Los sindicatos		,112	,324	***,778***
Los medios de comunicación	,318	,181		***,717***
Las organizaciones empresariales	,288	,391	,312	***,518***

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser. La rotación ha convergido en 6 iteraciones.

De hecho, la solución de cuatro factores revela cuatro ámbitos diferentes según los actores valorados. Una dimensión de poder "*directo*" o autoridad (Fuerzas de orden público, Fuerzas armadas, Monarquía, Iglesia católica), otro factor de poder "*judicial*" (Tribunales), la tercera de poder "*político*" en una democracia (gobierno, parlamentos y partidos) y el cuarto factor "*socioeconómico*" incluyendo el cuarto poder (los medios de comunicación).

Una de las conclusiones es que la valoración que la sociedad ha efectuado toma como criterio de referencia el "poder". Otra conclusión, que la división de tres poderes que propusiera Montesquieu, desde el punto de vista de la opinión pública española, aparece estructurado en la práctica en cuatro, con el poder legislativo y ejecutivo combinado en una dimensión de a) "representación", un poder referido a los responsables del b) orden (social y moral), otro factor c) poder "judicial" y un cuarto con los b) poderes económicos y de los medios de comunicación. En cierto sentido, implica el reconocimiento de los medios de comunicación como actores "económicos", empresas que buscan el beneficio económico sobre la información.

La agrupación de los actores sociales e instituciones, en este interés exploratorio, puede facilitar información especialmente relevante. Un ejemplo de ello es la valoración de la Iglesia Católica en Chile y Argentina tras los periodos de dictadura militar en la segunda mitad del siglo XX (Alaminos, 1987; 1991). La Iglesia Católica, en Chile, aparecía agrupada con la dimensión democrática: universidades, partidos de la oposición al general Pinochet, o medios de comunicación democráticos; por el contrario, en Argentina, la opinión pública ubicaba a la Iglesia Católica junto a la dictadura: el ejército, gobierno e instituciones antidemocráticas. Esta doble posición de la Iglesia Católica según el país viene explicada por el diferente papel que ejerció durante la represión militar. Mientras que la jerarquía de la iglesia católica chilena actuó de forma más protectora con los represaliados, la jerarquía argentina se alineó claramente con la dictadura, dándole legitimidad y apoyo. En el caso de España, donde la Iglesia Católica fue un factor esencial de legitimación y apoyo a la dictadura del general Franco, aún es percibida como un actor conservador e institucional.

Como podemos apreciar, las rotaciones permiten evaluar la percepción social de los actores, los criterios latentes para efectuar dichas evaluaciones, así como el modo en que forman agrupaciones con sentido teórico.

Como se debatió al inicio de este texto, una de las cuestiones centrales es nombrar las realidades que puedan detectarse mediante la exploración con el análisis factorial. Emplear un concepto u otro conduce a desarrollos argumentales muy diferentes, como puede apreciar el lector en los párrafos anteriores.

4.7. Las puntuaciones factoriales

Como resultado del procedimiento es posible crear una nueva variable para cada factor, donde se recojan las puntuaciones de cada caso en el factor. Recordemos que el procedimiento aspira a medir dimensiones que son expresadas de forma observable mediante indicadores y variables. El análisis factorial es, entre otras utilidades, un procedimiento para reducir la multivaria-

bilidad. En ese sentido, es interesante determinar la puntuación de cada caso en ese factor que está expresando y resumiendo un grupo de variables con significado teórico.

Una de las formulaciones matemáticas que hemos empleado expresa al factor o componente como resultado de una combinación lineal de las variables o indicadores empleados. Esa formulación es la que nos permite calcular la puntuación para cada caso en cada factor. La notación para estimar un componente o factor J, (F_j),

$$F_j = W_{j1}I_1 + W_{j2}I_2 + \ldots + W_{jn}I_n$$

donde las W son los coeficientes de puntuaciones factoriales, y n expresa el número de variables o indicadores (I).

Existen varios métodos para estimar la puntuación (valor) que cada caso muestra en cada factor. Estas puntuaciones, por lo general estandarizadas, pueden usarse en análisis posteriores como son regresión, análisis de varianza, discriminante y cualquier procedimiento que admita variables continuas.

Sin embargo, existen varias observaciones respecto a los métodos para generar dichas puntuaciones. Especialmente importante es la distinción entre el método de Componentes Principales y los demás métodos de extracción de factores empleados en el análisis factorial. Aunque el primero puede considerarse uno de los métodos disponibles para efectuar un análisis factorial, ya se comentó anteriormente que ambos proceden de modelos matemáticos y teóricos diferentes.

Los componentes principales son combinaciones lineales ponderadas de las variables observadas, mientras que los factores son variables latentes o no observadas, que se piensan son responsables de las correlaciones apreciadas entre variables observadas. En la práctica, implica que las puntuaciones que se generan mediante diferentes métodos (incluyendo Máxima Verosimilitud) al efectuar el análisis factorial están indeterminadas, mientras que en el modelo de componentes principales las puntuaciones son únicas.

En el caso del análisis factorial, con cualquier método de extracción que no sea componentes principales, no se dispone de suficiente información para estimar de forma única las puntuaciones. Independientemente del método de extracción o la realización o no de rotaciones, debido a que existen más parámetros no observados que datos observados. Este problema de la indeterminación de las puntuaciones en el análisis factorial ha sido bien estudiado y documentado[14].

14. Steiger (1979), Rozeboom (1982), Harman (1976), Mulaik (1972), Gnanadesikan (1977), Mardia, Kent, y Bibby (1979), Afifi, May, y Clark (2004), Clarkson y Jennrich (1988) o Dixon (1992).

Algunos programas facilitan una puntuación factorial estimada mediante regresión. Sin embargo, estas no pueden considerarse en puridad estadística una estimación como tal. En el caso de SYSTAD, solamente permite crear y guardar un factor con las puntuaciones cuando se utiliza el método de componentes principales. No permite la opción cuando se emplea cualquier otro método en el análisis factorial.

Otros programas, como SPSS, ofrecen tres métodos para estimar aproximadamente una puntuación para cada caso en cada factor. Sin embargo, cuando se emplee el método de componentes principales para la extracción, siempre generará el valor para cada caso en el factor con este método: factores ortogonales y puntuaciones estimadas con propiedades estadísticas adecuadas. El programa impone el criterio del método de componentes principales, independientemente de que se solicite algún tipo de rotación oblicua posterior o se seleccione otro procedimiento para calcular los valores en el factor. En ese sentido, actúa como SYSTAT cuando se emplean componentes principales.

Los métodos de regresión pueden emplearse también para estimar aproximadamente los valores de cada caso en el factor. Existen muchos métodos alternativos (Tucker, 1971; Harman, 1967) que producen puntuaciones con diferentes propiedades. En el caso de aplicar el método de regresión en SPSS (ajustando anteriormente con métodos diferentes a componentes principales), producen puntuaciones con una media de cero y varianza igual a la correlación múltiple al cuadrado entre las puntuaciones estimadas en los factores y los valores reales. Los factores pueden estar correlacionados incluso con una rotación ortogonal[15]. Otro método que ofrece SPSS es Bartlett (Bartlett 1937, 1938), donde nuevamente las puntuaciones estimadas tienen una media de cero. El procedimiento intenta minimizar la suma de cuadrados de los factores únicos. Por último, la tercera opción que permite el programa es Anderson-Rubin, que partiendo de una modificación del método anterior, estima puntuaciones con una media de cero, una desviación típica de 1 y los factores son ortogonales. Una información más detallada puede encontrarse en Harman (1976) o Lawley and Maxwell (1971). En ocasiones, la media de cero no es

15. Aunque la correlación entre dos factores se defina como cero (ortogonales) desde el plano teórico, el cálculo matricial no siempre permite esa situación. Equivale a cuando en una regresión de una variable "y" sobre una variable "x" se exige que el error "e_y" no esté correlacionado con la variable "y", ($y = a + bx + e_y$) (ecuación a). Si ese criterio se respeta, ya no puede ser satisfecho y cumplido en la regresión de "x" sobre "y", ($x = a + by + e_x$) (ecuación b). Las restricciones en una ecuación (a) condiciona a la otra (b) y pone en contradicción la exigencia teórica y la práctica.

exacta debido sobre todo a las consecuencias acumuladas del redondeo. En lo que se refiere a la desviación típica de 1, en ocasiones el método de estimación empleado no logra ajustar la desviación típica a dicho valor, excepto cuando el modelo factorial ajusta perfectamente[16].

Por lo general, si consideramos la producción investigadora internacional, lo más habitual en un análisis factorial es utilizar el método de componentes principales y la rotación ortogonal, lo que, hasta cierto punto, no deja de ser una hibridación entre modelos matemáticos y su referencia teórica.

El análisis factorial es un método muy interesante por su plasticidad y las facilidades que otorga al investigador para conectar conceptos teóricos con mediciones empíricas, ya sean variables o indicadores.

4.8. El análisis factorial para la construcción de índices

Esa posibilidad, de establecer mediante las correlaciones entre indicadores la existencia de un "factor", permite emplear los resultados como un índice que agrupe y combine, de forma multivariable, los diferentes indicadores. Por ejemplo, en el caso de determinar un índice que exprese el concepto democracia, este concepto podría se expresado por un factor que mantendría la siguiente relación con los indicadores.

$$\text{Indicador1} = a_1\ (\textit{Democracia}) + U_{ind1}$$
$$\text{Indicador 2} = a_2\ (\textit{Democracia}) + U_{ind2}$$
$$\text{Indicador 3} = a_3\ (\textit{Democracia}) + U_{ind3}$$
$$(...../.....)$$
$$\text{Indicador n} = a_n\ (\textit{Democracia}) + U_{indn}$$

Como ejemplo, consideremos el ejemplo del Índice de Desarrollo Humano (IDH). En la tabla se muestran varios países de América Latina y sus valores en el IDH. En este índice se consideran tres dimensiones: Educación, Salud y Estándar de vida. Estas tres dimensiones se miden con cuatro indicadores: "esperanza de vida al nacer", "media de años escolarizados", "años esperados de escolarización" y "PIB per cápita". En los análisis, se utilizarán los 187 países con datos y los indicadores.

---

16. Es exactamente la misma situación por la que cuando se efectúa una regresión de la variable "y" sobre la variable "x", ($y = a + bx + e_y$) (ecuación a). Si giramos la ecuación y hacemos la regresión de "x" sobre "y", ($x = a + by + e_x$) (ecuación b) no se logra los mismos valores en la ecuación (a) que en la (b), excepto cuando existe una colinealidad perfecta (y por lo tanto sobraría el error e).

IDH 2013. Indicadores para varios países

	IDH 2013	Esperanza de vida al nacer	Media de los años escolarizados	Años esperados de escolarización	PIB per cápita
41 Chile	0,822	80	9,8	15,1	20804
44 Cuba	0,815	79,3	10,2	14,5	19844
49 Argentina	0,808	76,3	9,8	16,4	17297
50 Uruguay	0,79	77,2	8,5	15,5	18108
65 Panamá	0,765	77,6	9,4	12,4	16379
67 Venezuela	0,764	74,6	8,6	14,2	17067
68 Costa Rica	0,763	79,9	8,4	13,5	13012
71 México	0,756	77,5	8,5	12,8	15854
79 Brasil	0,744	73,9	7,2	15,2	14275
82 Perú	0,737	74,8	9	13,1	11280
98 Colombia	0,711	74	7,1	13,2	11527
98 Ecuador	0,711	76,5	7,6	12,3	9998
102 República Dominicana	0,7	73,4	7,5	12,3	10844
111 Paraguay	0,676	72,3	7,7	11,9	7580
113 Bolivia	0,667	67,3	9,2	13,2	5552
115 El Salvador	0,662	72,6	6,5	12,1	7240
125 Guatemala	0,628	72,1	5,6	10,7	6866
132 Nicaragua	0,614	74,8	5,8	10,5	4266

Fuente: http://hdr.undp.org/es/data

Más adelante profundizaremos y explicaremos en detalle el significado de los resultados de un análisis factorial. Ahora, a modo introductorio, podemos afirmar que respecto a los indicadores anteriores, el análisis factorial muestra un único factor realmente significativo. Un factor que explica el 75% de la varianza de todos los indicadores.

Varianza total explicada de los indicadores de IDH

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3,032	75,798	75,798	3,032	75,798	75,798
2	,498	12,448	88,245			
3	,274	6,840	95,086			
4	,197	4,914	100,000			

Método de extracción: Análisis de Componentes principales

Además, todos los indicadores muestran una carga significativa en el índice estimado mediante la medición del factor.

Matriz de componentes

	Componente
	1
Esperanza de vida al nacer	,892
Media de años escolarizados	,894
Años esperados de escolarización	,911
(PIB) per cápita	,779

Método de extracción: Análisis de componentes principales

En ese sentido, el análisis factorial ofrece una forma multivariante de simplificar los indicadores utilizados y la realidad que expresan. Es en el cálculo del índice donde podemos adoptar estrategias alternativas. En el cálculo habitual del IDH, se emplean procedimientos aritméticos para integrar todos los valores (de cada indicador) en el índice final. El procedimiento del análisis factorial también produce una integración de los indicadores para producir un índice final. Sin embargo, el procedimiento será un poco más complejo y siguiendo una lógica de ajuste multivariable. En la tabla siguiente podemos comparar las puntuaciones Z del factor (es decir, centradas con una media de cero) con el índice del IDH calculado a partir de los mismos valores.

Comparación IDH con estimación multivariable y como índice compuesto

Posición	País	IDH 2013	Índice calculado como puntuación factorial
41	Chile	0,822	0,79656
44	Cuba	0,815	0,73501
49	Argentina	0,808	0,76182
50	Uruguay	0,79	0,58271
65	Panamá	0,765	0,3311
67	Venezuela	0,764	0,35384
68	Costa Rica	0,763	0,38072
71	México	0,756	0,27599
79	Brasil	0,744	0,26239
82	Perú	0,737	0,20223
98	Colombia	0,711	0,00646
98	Ecuador	0,711	0,02155
102	República Dominicana	0,7	-0,0796
111	Paraguay	0,676	-0,1847
113	Bolivia	0,667	-0,09759
115	El Salvador	0,662	-0,27395
125	Guatemala	0,628	-0,5305
132	Nicaragua	0,614	-0,4786

Fuente: elaboración propia sobre datos IDH

Este método de estimación puede dar resultados diferentes. En este caso, por ejemplo, cambian de posición países como Panamá (IDH 0,765), Venezuela (IDH 0,764) y Costa Rica (IDH 0,763), y que determinando el índice mediante análisis factorial (estadística multivariante) ofrece el orden inverso con Costa Rica (0,38), Venezuela (0,35) y Panamá (0,33). Muy posiblemente, en la medida que el peso de la "Educación" es más elevado en el índice estimado mediante análisis factorial. En el índice calculado mediante el análisis factorial, Ecuador aparece en una posición media para el conjunto de los países considerados. Su valor es de 0,02 cuando la media es cero. Evidentemente, puede normalizarse las puntuaciones factoriales por cualquiera de los procedimientos considerados en el capítulo 2.

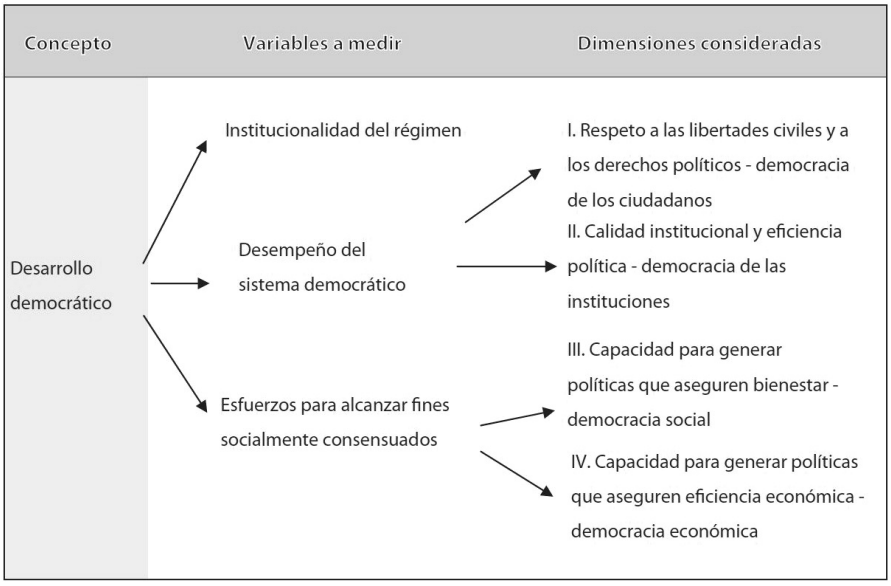
La conclusión es evidente. El análisis factorial (o de componentes principales) permite conocer el peso empírico de cada indicador en el índice final. En todo caso, facilita información sobre los pesos a utilizar en el caso de desear mantener un índice compuesto. Y, especialmente, confirma la posibilidad de medir una dimensión única mediante esos cuatro indicadores. Es decir, la potencia de la unidimensionalidad se impone sobre los rasgos particulares de cada dimensión.

También puede que aparezcan dimensiones que dificultan la creación de un único factor para todos los indicadores. Esta realidad habla de otras estructuras que contaminan la medición, oscurecen la imagen que ofrece un índice único y que deben diagnosticarse. Especialmente, dado que también influirían en cualquier otra estrategia para combinar los indicadores. Por lo tanto, como sabemos, la idea básica tras el análisis factorial es que pueden existir dimensiones latentes no visibles, que dan cuenta del comportamiento de grupos de variables visibles u observables directamente. Por lo tanto la finalidad del análisis factorial es identificar un conjunto de factores no observables que den cuenta de la correlación entre un conjunto de variables observadas. En cierto sentido, una de las importancias y ventajas del análisis factorial es que permite visualizar y revelar ruidos o sesgos que pueden quedar ocultos en la elaboración de índices complejos mediante operaciones de suma, resta, división, multiplicación. Es el caso cuando para calcular el índice se emplean índices que agrupan otras dimensiones.

Vamos a utilizar como ejemplo el Índice de Desarrollo Democrático. El concepto contempla cuatro dimensiones medidas por varios indicadores. Como se recoge en su metodología "El IDD-Lat se construye a partir de la agregación de varios indicadores que se ponderan, asignando puntajes tanto a las dimensiones como a cada uno de los indicadores seleccionados. En este punto importan dos cuestiones: a) la regla de agregación que se va a utilizar, y b) los pesos que se dan a las dimensiones que se agregarían y a sus componentes.

a) Regla de agregación utilizada: Multiplicativa. Los componentes de cada dimensión se ponderan entre 0 y 10, de tal forma que se cumpla que la sumatoria sea igual a 10.
b) Pesos que se asignan a las dimensiones y sus componentes: Se utilizan dos procedimientos para la asignación de puntajes a los indicadores, los que se distribuyen en una escala de 1 a 10. Es un programa de investigación muy importante y sólido, que busca armonizar y establecer unos criterios objetivos de referencia para medir la calidad democrática. En ese sentido, se utilizan sus datos para ejemplificar la dificultad de integración de heterogeneidad dentro de un único índice, tanto en la estrategia multivariante como agregativa.

Esquema 1. Estructura jerárquica de los conceptos. Análisis de la estructura lógica. Análisis multinivel



Fuente: http://www.idd-lat.org/2014/

Las dimensiones y sus valores por países fueron expuestas en la Tabla 1 en capítulos anteriores. Esas dimensiones son combinadas mediante operaciones aritméticas para calcular un índice que expresa la calidad de la democracia IDD-lat[17]. Vamos seguidamente a considerar en qué modo las cuatro dimensiones definen con consistencia un solo índice, mediante el análisis factorial. En el caso de definir dimensiones diferenciadas, su integración en un solo índice se convierte en un proceso de integración de heterogeneidad. El método utilizado en el análisis factorial es el de componentes principales.

Si tomamos como referencia el autovalor Eigen del primer componente o factor, obtenemos un solo índice estimado de forma multivariable que reflejaría un 58,4 de la varianza total. Podemos observar que aún queda un porcentaje elevado de varianza total por explicar. En términos de autovalor, una única dimensión resumiría la variabilidad de forma bastante limitada.

17. La metodología en detalle se puede consultar en http://www.idd-lat.org/2014/cuestiones_metodologicas/n/index.html

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,337	58,417	58,417	2,337	58,417	58,417
2	,924	23,088	81,504			
3	,483	12,079	93,584			
4	,257	6,416	100,000			

Método de extracción: Análisis de Componentes principales

En términos de mediciones, observamos que el emplear un solo índice elaborado de forma multivariable refleja de forma especial el peso de las dimensiones I y III. Las dimensiones II y IV aparecen menos representadas, o con menos influencia en la construcción del índice multivariable.

Matriz de componentes IDD-lat

	Componente
	1
DIM I	,834
DIM II	,588
DIM III	,863
DIM IV	,742

Método de extracción: Análisis de componentes principales

El segundo autovalor es de ,92 por lo que la solución de dos factores es una posibilidad muy plausible. Para dos factores, la varianza explicada es máxima, del 81,5%. Es decir, que la solución de agrupar las cuatro dimensiones en un solo índice significa sacrificar una heterogeneidad interna importante. La tabla siguiente muestra el resultado de mantener dos factores como solución de la estructura que muestran los datos.

Varianza total explicada IDD-lat con dos factores

Componente	Autovalores iniciales			**Sumas de las saturaciones al cuadrado de la extracción**			**Suma de las saturaciones al cuadrado de la rotación**		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,337	58,417	58,417	2,337	58,417	58,417	1,830	45,759	45,759
2	,924	23,088	81,504	,924	23,088	81,504	1,430	35,746	81,504
3	,483	12,079	93,584						
4	,257	6,416	100,000						

Método de extracción: Análisis de Componentes principales

Cuando evaluamos la carga de cada dimensión del índice IDD-lat sobre los dos factores, reconocemos una estructura que nos es conocida. La dimensión III y IV muestran una carga elevada en el factor 1, mientras que las dimensiones I y II tienen su carga más elevada en el factor 2. La dimensión I tiene su peso más repartido entre los dos factores, con una carga de ,523 en el factor I y de ,693 en el factor II. Esta solución nos recuerda la ofrecida por el análisis de clúster efectuado anteriormente, en el capítulo 2, con la que es plenamente consistente.

Matriz de componentes rotados IDD-lat con dos factores

	Componentes	
	Factor 1	Factor 2
DIM I	,523	,693
DIM II	,036	,934
DIM III	,873	,273
DIM IV	,891	,048

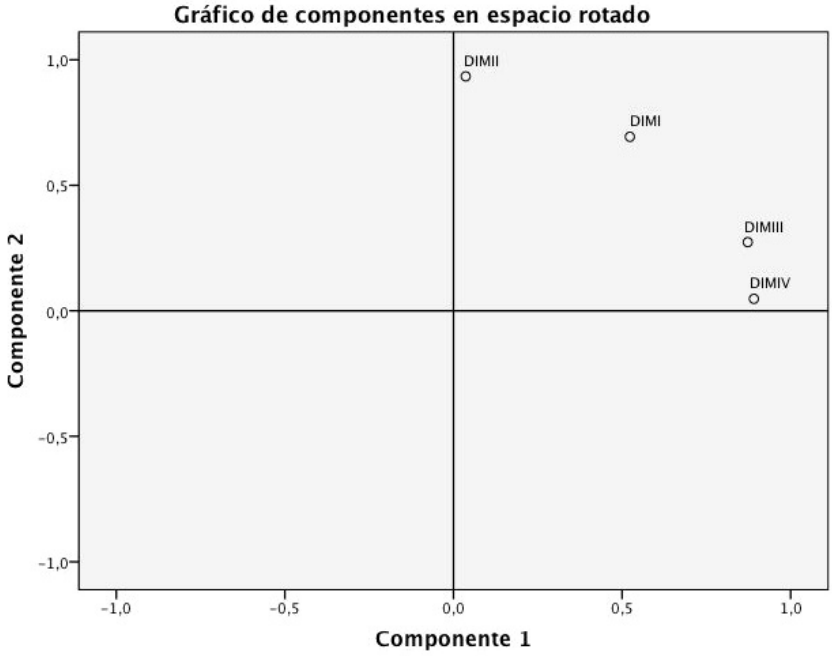
Método de extracción: Análisis de componentes principales. Método de rotación: Normalización Varimax con Kaiser. La rotación ha convergido en 3 iteraciones

Por eso, tal y como se recogía en el cuadro anterior, las siguientes dimensiones caracterizan la calidad de la democracia, tal y como son definidas por IDD-lat: "*Dimensión I: "Democracia de los ciudadanos*". Evalúa el respeto de los derechos políticos y las libertades civiles. *Dimensión II: "Democracia de las ins-*

tituciones". Mide la calidad institucional y la eficiencia del sistema político. *Dimensión III. "Democracia social y humana*". Analiza la capacidad del sistema democrático para generar políticas que aseguren bienestar y desarrollo humano. *Dimensión IV. "Democracia económica*". Expresa la capacidad para generar políticas que aseguren la eficiencia económica."

Basándonos en el análisis estadístico parece que el índice realmente contiene dos mediciones que deben ser armonizadas para poder considerarse un índice único de democracia. Las dimensiones III y IV parecen poseer un profundo carácter económico (políticas de bienestar y eficiencia económica), la dimensión II es de marcado acento político. La dimensión I se solapa lo "social" entre lo "económico" y lo "político". En ese sentido, existen evidentes dificultades de integración en un índice único (partiendo de los sistemas de subdimensiones e indicadores empleados) de las dimensiones social, política y económica que caracterizan la democracia.

El gráfico de componentes en el espacio rotado muestra como la dimensión II está muy próxima al eje del factor 1, y las dimensiones III y IV muy próximas al eje del factor 2. La dimensión I se encuentra próxima a la bisectriz del primer cuadrante (entre los dos ejes), indicando con ello que su carga está repartida entre los dos factores.

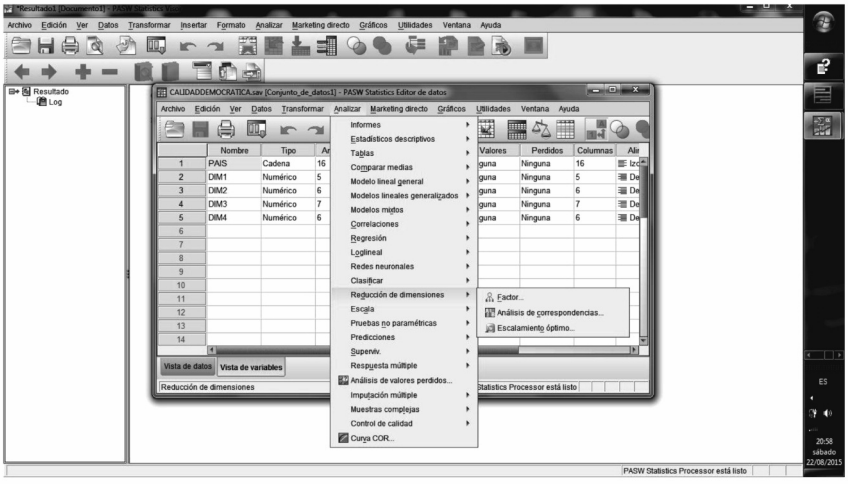


En ese sentido, una de las conclusiones del análisis factorial es la dificultad de sintetizar en un índice aritmético la variabilidad existente. De hecho, cuando consideramos las puntuaciones de los diferentes países en cada uno de los factores, observamos que son cuatro países los que despliegan el peso específico del factor 2, es decir de las dimensiones III y IV: Uruguay, Costa Rica y Chile (segundo cuadrante), y el Salvador (primer cuadrante). Cuando la relación entre los indicadores puede dar origen a varios factores, y no solamente a uno, expresa que realmente son varios los conceptos, o dimensiones de un concepto, los que se miden con ese grupo de indicadores. En definitiva, incluso un indicador puede ser polisémico y expresar varios conceptos parcialmente.

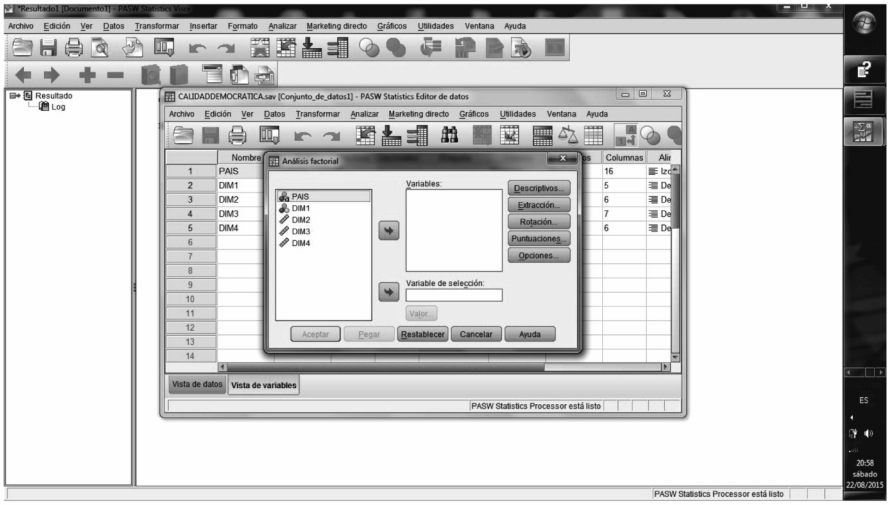
4.9. El análisis factorial con SPSS y SYSTAD

Los conceptos y procedimientos anteriores son ejecutados mediante programas estadísticos. Con algunos matices, prácticamente todos efectúan los mismos pasos, desde la selección de variables, opciones de rotación o métodos de extracción, alternativas para crear las variables (factores) o coeficientes de diagnóstico.

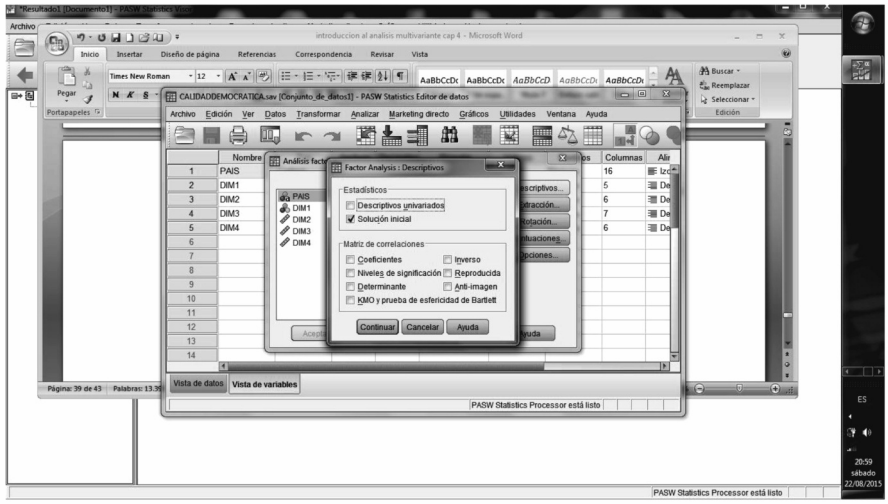
En el programa SPSS, la opción de análisis factorial se encuentra en el menú principal de "analizar" y la opción "reducción de dimensionalidad".



En la ventana principal se puede elegir las variables que formarán parte del análisis. En las situaciones en que se deseen filtrar los casos que van a considerase, la opción variable de selección permite hacerlo. Es, por ejemplo, que se desee efectuar el análisis para una categoría concreta en esa variable (por ejemplo, en variable género solamente para mujeres o en encuestas internacionales, elegir un país en concreto).



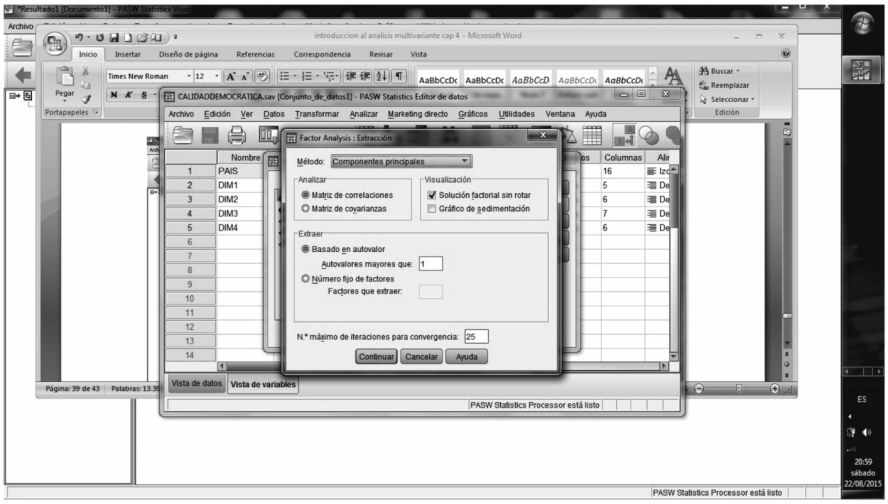
En la opción "Estadísticos" el procedimiento nos ofrecerá información sobre la media, la desviación estándar y el número de casos válidos en cada una de las variables que forman parte del análisis. La opción de mostrar "solución inicial" muestra la solución antes de rotación, incluyendo las comunalidades iniciales, los autovalores y el porcentaje de varianza explicada. En la opción "matriz de correlaciones" puede solicitarse los coeficientes de correlación, los niveles de significación, el determinante de la matriz, los coeficientes KMO y el test de esfericidad de Bartlett, así como las matrices inversa, reproducida y la AIC.



La opción "Extracción" permite que se indique un método de extracción. El programa dispone de varios métodos como son componentes principales, mínimos cuadrados no ponderados, mínimos cuadrados generalizados, máxima verosimilitud, etc.

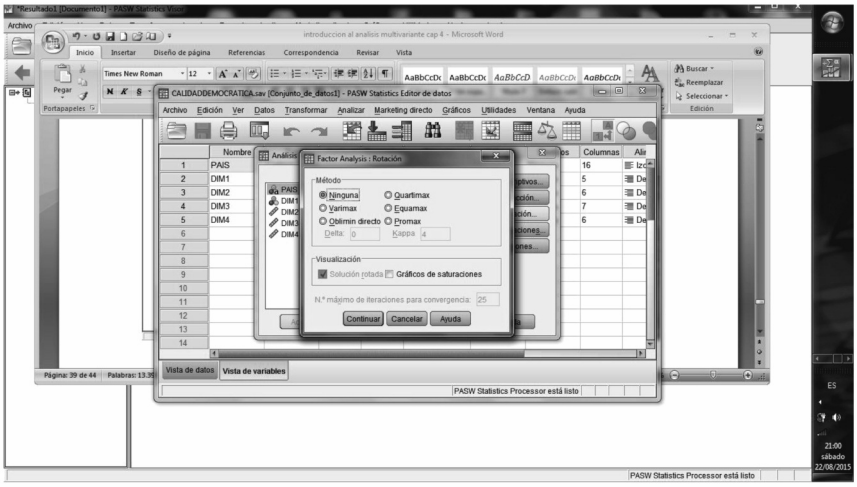
Es posible, asimismo, determinar si se desea utilizar para el análisis la matriz de correlación o la matriz de covarianzas. La matriz de correlaciones es útil cuando las variables están medidas en diferentes escalas, mientras que la matriz de covarianzas es preferible cuando el análisis se va a aplicar en múltiples grupos con diferentes varianzas en las variables consideradas.

Con la opción de extracción se decide si se retienen en el análisis los factores con un valor eigen igual o superior a 1, o por el contrario se desea indicar cuantos factores se desean mantener en el análisis. También se ofrece la posibilidad de mostrar la solución factorial no rotada y el gráfico de sedimentación. Finalmente, esta ventana ofrece la opción de decidir el número máximo de iteraciones que puede emplear el algoritmo para alcanzar una solución.



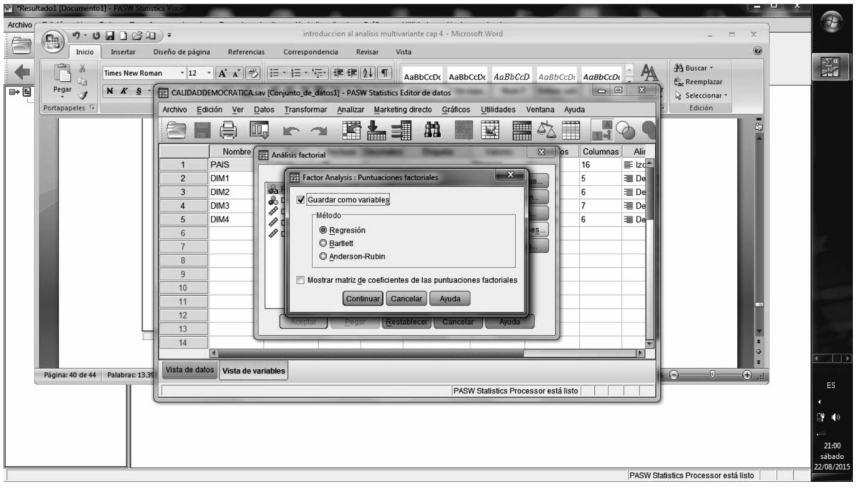
La opción "Rotación" permite elegir el método de rotación que se desee ejecutar, sea ortogonal o no. En el caso de SPSS las rotaciones disponibles son varimax, direct oblimin, quartimax, equamax, o promax. El gráfico de componentes muestra mediante gráficos bivariables, la relación de las variables o indicadores con los factores. Al igual que con la extracción, es posible indicar un número máximo de iteraciones para efectuar la rotación.

Las puntuaciones factoriales se pueden guardar en variables mediante la opción "guardar como variables". Como consecuencia de seleccionar dicha opción se creará tantas variables como factores se conserven en el análisis, y en cada una de ellas se guardará la puntuación de cada caso en cada factor. Los méto-



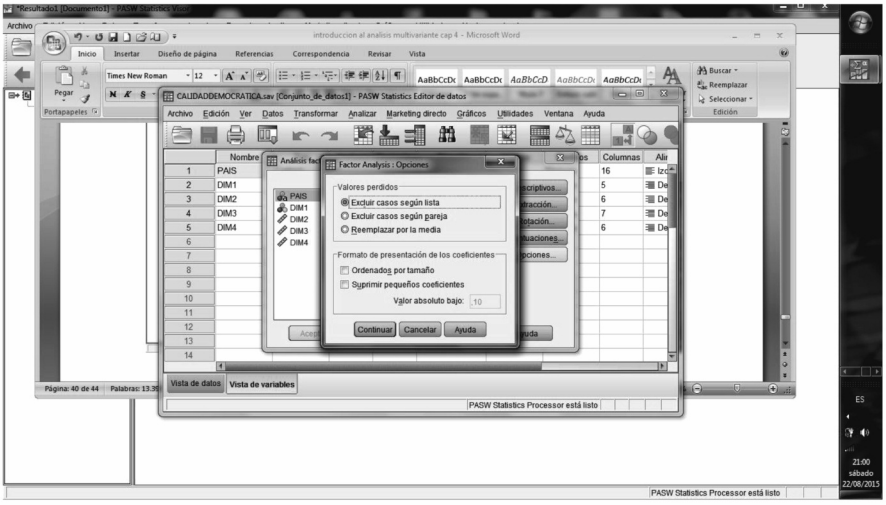
dos utilizados son estimaciones mediante regresión, Bartlett, y Anderson-Rubin. Recordar que cuando se emplea el método de componentes principales, los factores que se guarden serán ortogonales independientemente de que se rote oblicuo o se pida otro método para estimar las puntuaciones.

La matriz de coeficientes factoriales muestra los valores que relacionan variables y factores, para estimar las puntuaciones. Equivalen a los coeficientes en la ecuación de regresión múltiple que expresa a cada factor como una combinación lineal de variables o indicadores. Las variables se multiplican por dichos coeficientes para estimar la puntuación de cada caso en cada factor. Así mismo, también ofrece la matriz de correlación entre los factores.

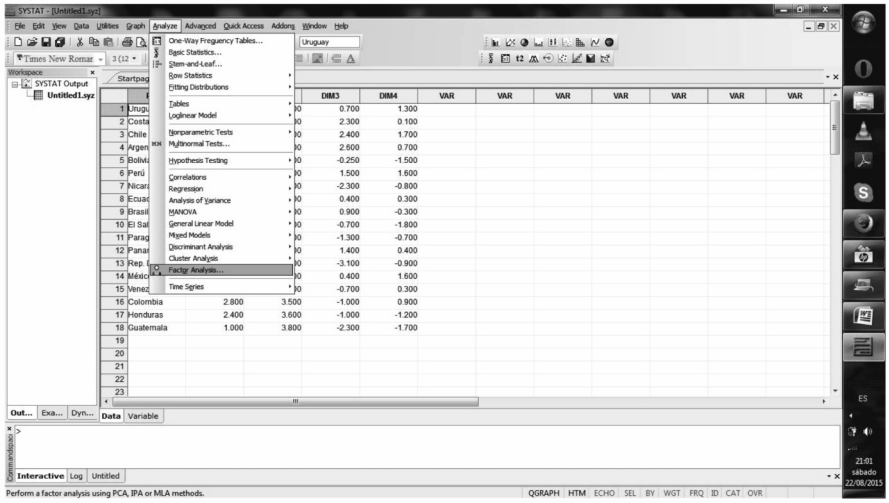


Por último, en las "opciones" puede decidirse el tratamiento que se dará a los casos perdidos, con las opciones de excluir los casos por parejas, eliminar los casos con algún valor perdido en cualquiera de las variables consideradas, o sustituir los valores perdidos por el valor medio de cada variable. En el caso del borrado por pares cada coeficiente de correlación en la matriz estará calculado sobre un tamaño muestral diferente. El borrado por lista tiene como consecuencia una reducción importante del tamaño muestral. La sustitución por la media introduce una "suavización" en las distribuciones que afectará a las cargas de variables en los factores. En todo caso, las consecuencias del tratamiento de los casos perdidos deben ser siempre valoradas, en la medida que pueden condicionar las soluciones factoriales.

La última utilidad en opciones es la posibilidad de que las variables aparezcan ordenadas según su carga en cada factor en las matrices factoriales (sean de estructura o configuración). Esto facilita mucho la interpretación de los factores, dado que agrupa las variables que están más próximas entre ellas. En ese proceso de facilitar la lectura e interpretación de los resultados, pueden suprimirse del resultado los coeficientes excesivamente bajos. En ese caso, debe indicarse un valor de referencia para que no se muestre en el resultado.



El programa SYSTAD muestra directamente la opción para el análisis factorial en el menú "Analizar". En conjunto, tiene un diseño más compacto, donde las diferentes elecciones que deben efectuarse se encuentran agrupadas en pocas ventanas.



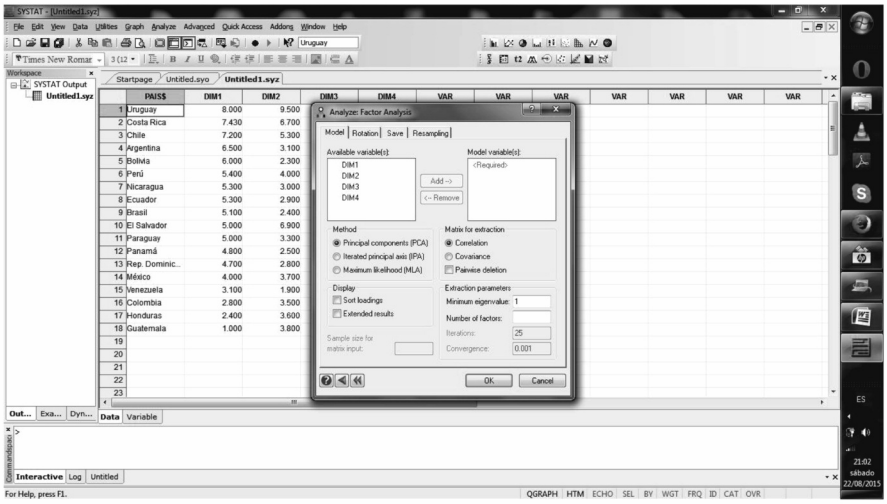
Así, en la ventana modelos aparece la opción de elegir las variables que se analizarán. En método de extracción aparece en primer lugar la opción más frecuente, de Componentes Principales. Otros dos métodos disponibles son Ejes Principales y Máxima Verosimilitud.

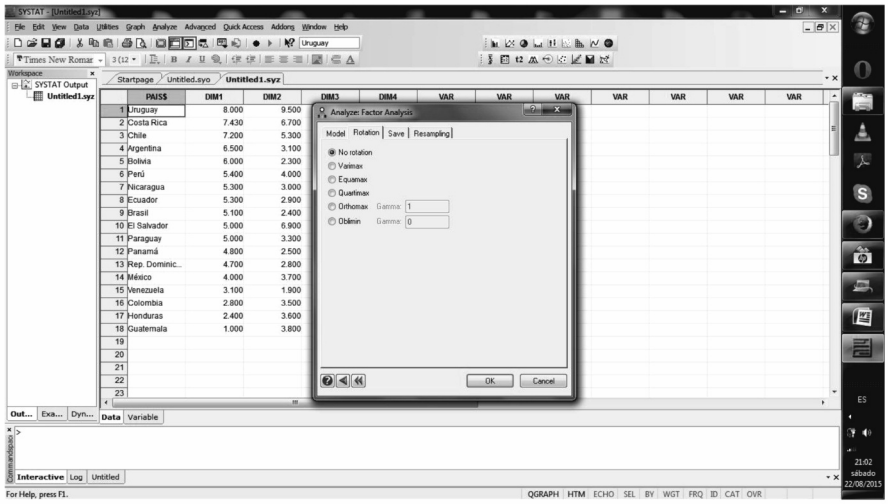
Se puede seleccionar para la extracción la matriz de correlaciones (variables con diferentes escalas) o la matriz de covarianzas (comparación de modelos entre diferentes poblaciones). Junto a esas opciones aparece el tratamiento de los casos perdidos, con la opción de exclusión por pares.

Puede elegirse el criterio para decidir el número de factores, ya sea indicando el valor mínimo que debe alcanzar el autovalor (eigen) de un factor para conservarse en la solución. Por defecto ofrece el valor 1. La otra opción es decidir cuántos valores se desea conservar en el análisis. Bajo dichas opciones se encuentra el número máximo de iteraciones para la extracción y como opción, el valor de convergencia para la solución. En esta pantalla se pueden decidir los criterios para mejorar la interpretación del resultado, por ejemplo ordenando las variables según la carga en cada factor.

En la pantalla de rotación se elige el tipo de rotación que se desea. Las rotaciones disponibles en SYSTAT son varimax, oblimin, quartimax, equamax, u ortomax. En el caso de ser oblicua, se introduce un valor gamma que controla el grado de asociación que admitimos en los factores, tal y como se explicó anteriormente.

Este programa ofrece la posibilidad de guardar bastante información como resultado del análisis. No obstante, para el caso de las puntuaciones factoriales emplea exclusivamente el método de componentes principales. Eso hace



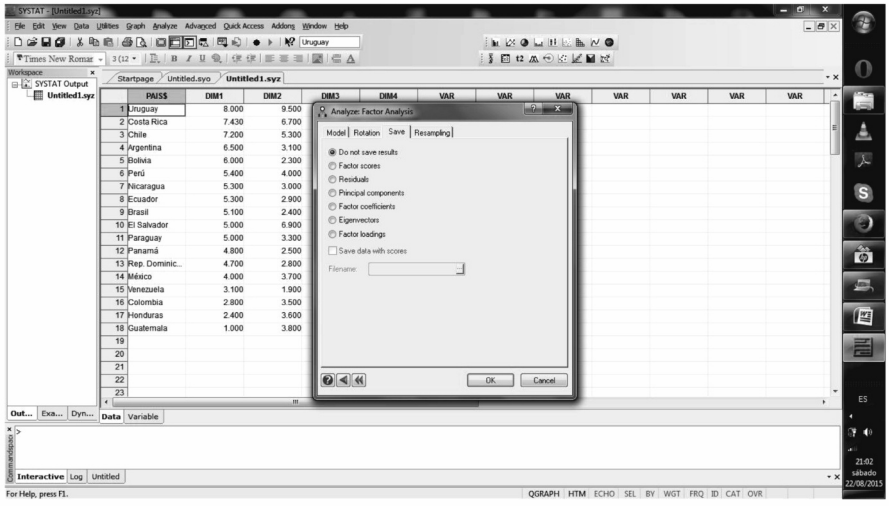


que para los métodos de máxima verosimilitud y ejes principales solamente pueden guardarse las cargas factoriales, pero no se estima puntuación alguna. De hecho, las opciones se desactivan de elegir algún método diferente a componentes principales.

Con el método de extracción de componentes principales es posible guardar las puntuaciones factoriales estandarizadas, los residuales para cada caso, las puntuaciones sin estandarizar de los componentes principales (solamente en extracción de componentes principales y sin rotación), los coeficientes fac-

toriales con los valores que relacionan variables y factores, para estimar las puntuaciones, los vectores eigen y las cargas factoriales. Otros programas ofrecen varias de estas posibilidades como parte de la información que se muestra con los resultados, mientras que este programa permite guardar los datos en archivo.

Una opción interesante es la de guardar las puntuaciones (no las cargas factoriales y otros datos) como variables (junto a las del archivo original) en un nuevo archivo tal y como vimos en el SPSS. Los factores se nombrarán de forma sucesiva (Factor (1), Factor (2)…), y aquellos casos con valores perdidos en alguna de las variables empleadas para el análisis factorial también tendrá valor perdido en el factor. Si se emplea una matriz de correlaciones los factores estarán estandarizados con media cero y varianza 1. Cuando se utiliza la matriz de covarianzas y no se efectúa rotación, las puntuaciones no estarán estandarizadas.



La última opción, muy presente en SYSTAT, es la posibilidad de testar, mediante muestreos de los datos, la fiabilidad del modelo que estamos empleando. Es un procedimiento muy interesante que excede los objetivos de este texto.

BIBLIOGRAFÍA

Afifi, A. A., May, S., and Clark, V. (2004). *Computer-aided multivariate analysis,* 4th ed. New York: Chapman & Hall.

Akaike, H. 1987. Factor analysis and AIC. *Psychometrika* 52: 317–332.

Alaminos, A.F. (1987) *Cultura política y económica en el cono sur: Argentina, Chile y Uruguay*. Madrid: CEDEAL.

Alaminos, A.F. (1991) *Chile: transición política y sociedad*. Madrid: Siglo XXI-CIS.

Alaminos, A.F. (1998). *Teoría y práctica de la encuesta. Aplicación a los países en vías de desarrollo*. Alicante: Club Universitario.

Alaminos, A.F. (2004). "Tendencias en ideología política: estructura y contenidos", en Tezanos, J.F. *Tendencias en identidades, valores y creencias*. Madrid: Sistema.

Alaminos, A.F. (2005). *El análisis de la realidad social. Modelos estructurales de covarianzas*. Alicante: OBETS.

Alaminos, A.F. (2005). *Introducción a la Sociología Matemática*. Alicante: SPES.

Anderberg, M. R. 1973. Cluster Analysis for Applications. New York: Academic Press.

Bacher, J. (1996). *Clusteranalyse: Anwendungsorientierte Einführung*. München: Oldenbourg. 2., ergänzte Auflage.

Bacher, J. (2000). A Probabilistic Clustering Model for Variables of Mixed Type. *Quality &Quantity*, 34, 223–235.

Bacher, J. (2002). StatistischesMatching: Anwendungsm¨oglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS. *ZA-Informationen*, 51, 38–66.

Bartlett, M. S. 1937. The statistical conception of mental factors. *British Journal of Psychology* 28: 97–104.

Bartlett, M. S. 1938. Methods of estimating mental factors. *Nature*, London 141: 609–610.

Bartlett, M. S. 1951. The effect of standardization on a 2 approximation in factor analysis. *Biometrika* 38: 337–344.

Basilevsky, A. T. 1994. *Statistical Factor Analysis and Related Methods: Theory and Applications*. New York: Wiley.

Bender, S., Brand, R., & Bacher, J. (2001). Re-identifying register data by survey data: An empirical study. *Statistical Journal of the UN Economic Commission for Europe*, 18(4), 373–381.

Bezdek, J.C and Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Trans. Systems, Man and Cybernetics*, *Part B: Cybernetics*, 28, 301-315.

Blashfield, R. K., and M. S. Aldenderfer. 1978. The literature on cluster analysis. *Multivariate Behavioral Research* 13: 271–295.

Bollen, K. A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

Calinski, T., and J. Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics* 3: 1–27.

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.

Cattell, R. B. 1966. The scree test for the number of factors. *Multivariate Behavioral Research* 1: 245–276.

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2001* (pp. 263–268).

Clarke, M. R. B. 1970. A rapidly convergent method for maximum-likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology* 23: 43–52.

Clarkson, D. B. and Jennrich, R. I. (1988). Quartic rotation criteria and algorithms. *Psychometrika*, 53, 251–259.

Day, W. H. E., and H. Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods.Journal of Classification 1: 7–24.

Davies, D.L. and Bouldin, D.W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell.*, 1, 4, 224-227.

Dixon, W. J. (1992). *BMDP statistical software manual*. Berkeley: University of California Press.

Duda, R. O., P. E. Hart, and D. G. Stork. 2001. Pattern Classification. 2nd ed. New York: Wiley.

Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *Journal of Cybernetics*, 3, 32-57.

Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.

Harman, H. H. 1976. *Modern Factor Analysis*. 3rd ed. Chicago: University of Chicago Press.

Horst, P. 1965. *Factor Analysis of Data Matrices*. New York: Holt, Rinehart & Winston.

Everitt, B. S. 1993.Cluster Analysis. 3rd ed. London: Arnold.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl. 2011.*Cluster Analysis*. 5th ed. Chichester, UK: Wiley.

Fisher, L. and Van Ness, J. W. (1971). Admissible clustering procedures. *Biometrika*, 58, 91–104.

Fox, J. (1982). *Selectiv Aspects of measuring Resemblance for Taxonomy*, (pp. 127–151). Jossey-Bass: San Francisco, Washington, London.

Fraley, C. & Raftery, A. E. (1998). How many Clusters? Which Clustering Method? Answers via Model-based Cluster Analysis. *Computer Journal*, 4, 578–588.

Francés, F.; Alaminos, A.; Penalva, C. y Santacreu, O. (2014). El proceso de medición de la realidad social: La investigación a través de encuestas. Cuenca: PYDLOS.

Fuller, W. A. 1987. *Measurement Error Models*. New York: Wiley.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. 2nd ed. New York: John Wiley & Sons.

Gordon, A. D. 1999. Classification. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Gorsuch, R. L. 1983. *Factor Analysis*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

Gower, J. C. (1967). A comparison of some methods of cluster analysis. *Biometrics,* 23, 623–637.

Gruvaeus, G. and Wainer, H. (1972). Two additions to hierarchical cluster analysis. *The British Journal of Mathematical and Statistical Psychology*, 25, 200–206.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 139–150.

Hamilton, L. C. 2009. *Statistics with Stata* (Updated for Version 10). Belmont, CA: Brooks/Cole.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: John Wiley & Sons.

Hartigan, J.A. and Wong, M. A. (1979), A K-Means Clustering Algorithm. *Applied Statistics*, 28, 100-108.

Harman, H. H. (1976). *Modern factor analysis*, 3rd ed. Chicago: University of Chicago Press.

Jackson, J. E. (2003). *A user's guide to principal components*. New York: Wiley Interscience.

Jain, A. K., and R. C. Dubes. 1988. Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice Hall.

Jennrich, R.I. and Robinson, S.M. (1969). A Newton-Raphson algorithm for maximum likelihood factor analysis. *Psychometrika, 34, 111-123*.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika,* 32, 241–254.

Jöreskog, K. G., and D. Sörbom. 1986. *Lisrel VI: Analysis of linear structural relationships by the method of maximum likelihood*. Mooresville, IN: Scientific Software.

Jöreskog, K. G., and D. Sörbom. 1988. *PRELIS: A program for multivariate data screening and data summarization. A preprocessor for LISREL*. 2nd ed. Mooresville, IN: Scientific Software.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27, 857–872.

Kaiser, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23: 187–200.

Kaiser, H. F. 1974. An index of factor simplicity. *Psychometrika* 39: 31–36.

Kim, J. O., and C. W. Mueller. 1978. *Introduction to factor analysis. What it is and how to do it*. In Sage University Paper Series on Quantitative Applications the Social Sciences, vol. 07–013. Thousand Oaks, CA: Sage.

Kim, J. O., and C. W. Mueller. 1978. *Factor analysis: Statistical methods and practical issues*. In Sage University Paper Series on Quantitative Applications the Social Sciences, vol. 07–014. Thousand Oaks, CA: Sage.

Lawley, D. N., and A. E. Maxwell. 1971. *Factor Analysis as a Statistical Method*. 2nd ed. London: Butterworths.

Holm, K. (2004). ALMO Statistik-System, Version 7.1. http://www.almo-statistik.de/.

Huang, Z. (1998). Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Variables. *Data Mining and Knowledge Discovery*, 2, 283–304.

Kaufman, L., and P. J. Rousseeuw. 1990.Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley.

Lance, G. N., and W. T. Williams. 1967. A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Computer Journal* 9: 373–380.

Lawley, D. N., and A. E. Maxwell. 1971. *Factor Analysis as a Statistical Method*. 2nd ed. London: Butterworths.

Ling, R. F. (1973). A computer generated aid for cluster analysis. *Communications of the ACM,* 16, 355–361.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *5th Berkeley symposium on mathematics, statistics, and probability*, 1, 281–298.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press.

McQuitty, L. L. (1960). Hierarchical syndrome analysis. *Educational and Psychological Measurement*, 20, 293–303.

Milan, L., and J. Whittaker. 1995. Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics* 44: 31–49.

Milligan, G. W. (1980). An examination of the effects of six types of error perturbation on fifteen clustering algorithms. *Psychometrika,* 45, 325–342.

Milligan, G. W., and M. C. Cooper. 1985. An examination of procedures for determining the number of clusters in a dataset. *Psychometrika* 50: 159–179 12

Milligan, G.W. (1987), A study of beta-flexible clustering method, *College of Administrative Science Working Paper Series*, 87-61 Columbus, OH: The Ohio State University.

Milligan, G. W., and M. C. Cooper. Introduction to cluster-analysis commands. 1988. A study of standardization of variables in cluster analysis. *Journal of Classification* 5: 181–204.

Morrison, D. F. (2004). *Multivariate statistical methods*, 5th ed. CA: Duxbury Press.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Penalva, C.; Alaminos, A.; Francés, F y Santacreu, O. (2015). *La investigación cualitativa: técnicas de investigación y análisis con Atlas. Ti*. Cuenca: PYDLOS

Preparata, G. and Shamos, M. (1985). *Computational geometry: An introduction*. New York: Springer-Verlag.

Raciborski, R. 2009. Graphical representation of multivariate data using Chernoff faces. *Stata Journal* 9: 374–387.

Rao, C. R. 1955. Estimation and tests of significance in factor analysis. *Psychometrika* 20: 93–111.

Rencher, A. C. 1998. *Multivariate Statistical Inference and Applications*. New York: Wiley.

Rencher, A. C., and W. F. Christensen. 2012. Methods of Multivariate Analysis. 3rd ed. Hoboken, NJ: Wiley.

Rohlf, F. J. 1982. Single-link clustering algorithms. In Vol. 2 of Handbook of Statistics, ed. P. R. Krishnaiah and L. N. Kanal, 267–284. Amsterdam: North-Holland.

Rost, J. (1985). A latent class model for rating data. *Psychometrika*, 50(1), 37–49.

SAS Institute Inc. (2002). *SAS OnlineDoc*. Cary, NC. http://v9doc.sas.com/sasdoc/

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired associate learning tasks. *Journal of Experimental Psychology*, 53, 94-101.

Rozeboom, W. W. (1982). The determinacy of common factors in large item domains. *Psychometrika*, 47, 281–295.

Sattath, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika*, 42, 319–345.

Schaffer, C. M., and P. E. Green. 1996. An empirical comparison of variable standardization methods in cluster analysis. *Multivariate Behavioral Research* 31: 149–167.

Sharma, S.C. (1995). *Applied multivariate techniques*. New York: John Wiley & Sons.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Sibson, R. 1973. SLINK: An optimally efficient algorithm for the single-link cluster method. *Computer Journal* 16: 30–34.

Silverman, B.W. (1986), *Density estimation*, New York: Chapman & Hall.

Snedecor, G.W. y Cochran, W.G. (1967) Statistical methods. Ames, Iowa: Iowa State University Press.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin,* 38, 1409–1438.

Sokal, R. R. and Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. San Francisco: W. H. Freeman and Company.

Spath, H. 1980. Cluster Analysis Algorithms for Data Reduction and Classification of Objects. Chichester, UK: Ellis Horwood.

Spearman, C. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15: 72–101.

SPSS Inc. (2001). *The SPSS TwoStep cluster component. A scalable component to segment your customers more effectively*. White paper – technical report, Chicago. ftp://ftp.spss.com/pub/web/wp/TSCWP-0101.pdf

SPSS Inc. (2004). *TwoStep Cluster Analysis*. Technical report, Chicago. http://support.spss.com/tech/stat/Algorithms/12.0/twostep cluster.pdf

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and 1970's: some interesting parallels. *Psychometrika*, 44, 157–167.

Tarlov, A. R., J. E. Ware Jr., S. Greenfield, E. C. Nelson, E. Perrin, and M. Zubkoff. 1989. The medical outcomes study. An application of methods for monitoring the results of medical care. *Journal of the American Medical Association* 262: 925–930.

Thomson, G. H. 1951. *The Factorial Analysis of Human Ability*. London: University of London Press.

van Belle, G., L. D. Fisher, P. J. Heagerty, and T. S. Lumley. 2004. *Biostatistics: A Methodology for the Health Sciences*. 2nd ed. New York: Wiley.

Vermunt, J. & Magidson, J. (2000). *Latent GOLD 2.0. User's Guide*. Belmont.

Vizirgiannis, M., Haldiki, M. and Gunopulos, D. (2003). *Uncertainity handling and quality assessment in data mining*. London: Springer-Varlag.

Wainer, H. and Schacht, S. (1978). Gappint. *Psychometrika,* 43, 203–212.

Ward, J. H., Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58: 236–244.

Wilkinson, L. (1979). Permuting a matrix to a simple structure. *Proceedings of the American Statistical Association,* 409–412.

Winer B.J. (1971) Statistical Principles in Experimental Design. New York: McGraw-Hill

Wishart, D. (2003). k-Means Clustering with Outlier Detection, Mixed Variables and Missing Values. In M. Schwaiger & O. Opitz (Eds.), *Exploratory data analysis in empirical research. Proceedings of the 25th Annual Conference of the Gesellschaft f¨ur Klassifikation e.V., University of Munich, March 14-16, 2001*, Studies in Classification, Data Analysis, and Knowledge Organization (pp. 216–226). Berlin: Springer.

Wong, M.A. and Lane, T. (1983), A kth nearest neighbor clustering procedure, *Journal of Royal Statistical Society*, Series B, 45 362-368.